Check for updates

Empirical Article

Advances in Methods and Practices in Psychological Science
April-June 2023, Vol. 6, No. 2,
pp. 1–23
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/25152459221147250
www.psychologicalscience.org/AMPPS

iCatcher+: Robust and Automated Annotation of Infants' and Young Children's Gaze Behavior From Videos Collected in Laboratory, Field, and Online Studies

Yotam Erel[1], Katherine Adams Shannon[2], Junyi Chu[3], Kim Scott[3], Melissa Kline Struhl[3], Peng Cao[4], Xincheng Tan[5], Peter Hart[3], Gal Raz[3], Sabrina Piccolo[3], Catherine Mei[3], Christine Potter[6], Sagi Jaffe-Dax[7], Casey Lew-Williams[8], Joshua Tenenbaum[3,9], Katherine Fairchild[9], Amit Bermano[1], and Shari Liu[3,10]

[1]The Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv-Yafo, Israel; [2]Department of Psychology, Stanford University, Palo Alto, California; [3]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts; [4]Computer Science & Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts; [5]School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts; [6]Department of Psychology, The University of Texas at El Paso, El Paso, Texas; [7]The School of Psychological Sciences, Tel Aviv University, Tel Aviv-Yafo, Israel; [8]Department of Psychology, Princeton University, Princeton, New Jersey; [9]The MIT Quest for Intelligence, Massachusetts Institute of Technology, Cambridge, Massachusetts; and [10]Department of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, Maryland

Abstract

Technological advances in psychological research have enabled large-scale studies of human behavior and streamlined pipelines for automatic processing of data. However, studies of infants and children have not fully reaped these benefits because the behaviors of interest, such as gaze duration and direction, still have to be extracted from video through a laborious process of manual annotation, even when these data are collected online. Recent advances in computer vision raise the possibility of automated annotation of these video data. In this article, we built on a system for automatic gaze annotation in young children, iCatcher, by engineering improvements and then training and testing the system (referred to hereafter as iCatcher+) on three data sets with substantial video and participant variability (214 videos collected in U.S. lab and field sites, 143 videos collected in Senegal field sites, and 265 videos collected via webcams in homes; participant age range = 4 months–3.5 years). When trained on each of these data sets, iCatcher+ performed with near human-level accuracy on held-out videos on distinguishing "LEFT" versus "RIGHT" and "ON" versus "OFF" looking behavior across all data sets. This high performance was achieved at the level of individual frames, experimental trials, and study videos; held across participant demographics (e.g., age, race/ethnicity), participant behavior (e.g., movement, head position), and video characteristics (e.g., luminance); and generalized to a fourth, entirely held-out online data set. We close by discussing next steps required to fully automate the life cycle of online infant and child behavioral studies, representing a key step toward enabling robust and high-throughput developmental research.

Keywords

cognitive development, online data collection, eye tracking, open source, deep learning

Received 5/5/22; Revision accepted 11/4/22

Corresponding Authors:
Yotam Erel, The Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv-Yafo, Israel
Email: erelyotam@gmail.com

Shari Liu, Department of of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, Maryland
Email: shariliu@jhu.edu

Creative Commons NonCommercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/), which permits noncommercial use, reproduction, and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

Where infants look, and for how long, has served as a primary outcome measure for developmental psychology since the birth of the field (Friedman, 1972; Haith, 1980; Horowitz et al., 1972; Slater et al., 1984; Teller, 1979). Experiments measuring infants' looking behavior have delivered insights into the developmental origins and nature of perception (Aslin & Smith, 1988), learning (Kirkham et al., 2002; Saffran et al., 1996), categorization (Waxman & Markow, 1995; Xu et al., 1999), preference for stimuli such as faces (Simion et al., 2008; Valenza et al., 1996), language processing (Lew-Williams & Fernald, 2007; Lukyanenko & Fisher, 2016), and understanding of domains such as people, objects, and number (Baillargeon et al., 2016; Feigenson et al., 2004; Hamlin et al., 2007; Spelke et al., 1992). Yet discovery in the field is constrained by two key bottlenecks that slow the pace of empirical research and limit its robustness and generalizability. The first obstacle is recruiting and testing adequately powered samples of infants (Byers-Heinlein et al., 2022; Frank et al., 2017; Oakes, 2017). Online platforms such as Lookit (Scott & Schulz, 2017) have been developed to allow families to participate in studies online via webcam, which enables faster and more efficient data collection, potentially in a much more diverse population than ever before.

Nevertheless, even with rapid data collection, a second obstacle still looms large: annotating video data from infants to produce outcome measures such as duration and direction of gaze. Roughly speaking, it takes an experienced human annotator 2 to more than 10 times as long as the duration of a video to generate outcome labels for that video, depending on the complexity and resolution of the measures and the characteristics of the video (e.g., movement, lighting). In this article, we expand on a promising system designed specifically for classification of young children's looking behavior, iCatcher (Erel et al., 2022; but see also Chouinard et al., 2019 and Werchan et al., 2022). We demonstrate its suitability for use in developmental research by (a) engineering technical improvements to iCatcher to extract accurate and robust frame-by-frame labels of gaze behavior from large video data sets of infants and toddlers in variable environments and (b) showing that the system's performance parallels the reliability of "gold-standard" manual annotation.

iCatcher: Solving the Gaze-Annotation Bottleneck

In the last decade, new tools in computer vision have enabled the estimation of gaze behavior given webcam videos, including OpenFace (Baltrusaitis et al., 2018), RT-GENE (Fischer et al., 2018), WebGazer (Papoutsaki et al., 2016), and Opengazer (Zielinski, 2007). These tools rely on extracting eye features and facial landmarks (e.g., eyes, nose, mouth) from video, which are then passed to deep-learning models to predict gaze direction. However, these approaches have been developed for relatively still adult faces, not squirming infants, and require high-quality video data, a condition that is often not met in online developmental studies (cf. Werchan et al., 2022). They also often require some manual labor and/or show reduced performance when videos contain variation in superficial features such as lighting conditions. Erel and colleagues (2022) improved these solutions by creating an openly available program, iCatcher, a neural-network approach rooted in computer-vision methods and specifically designed for the needs of research with infants and young children. iCatcher showed higher accuracy in estimating real-time gaze location relative to prior approaches by applying one key insight: Successive video frames are not independent from one another. iCatcher uses a moving window of five frames to estimate the gaze direction of the center frame (LEFT, RIGHT, or AWAY) and does so iteratively throughout a video recording of a child's face. This feature of the network architecture (among others) allows the neural network to be trained to classify eye gaze in a set of participant videos with somewhat higher accuracy than RT-GENE and dramatically higher accuracy than OpenFace.

Open Questions About Accuracy and Robustness to Video, Participant, and Experiment Variability

For iCatcher to become a viable tool for studies of cognitive development, particularly as the field is moving toward online data collection from more representative participant samples, iCatcher must be accurate (performing with near-human accuracy), robust (accurate over sources of video, experiment, and participant variability), and usable (accessible to all researchers in the field). Here, we focus on accuracy and robustness, leaving the challenge of usability for future work.

First, iCatcher should be able to support studies of infants and children tested in the lab, in the field, and at home (Tsuji et al., 2021). Erel et al. (2022) showed that iCatcher delivered human-level performance in one video data set, drawn from one lab, in which all participants were tested in the same setup (holding viewing distance, screen size, camera position, lighting, and backdrop constant). However, labs vary substantially in their methods and setups for in-person testing. Online testing introduces even more variability (for examples of still frames from webcam videos, see Fig. 1c). Ideally, iCatcher could be used to support online research on the Lookit platform (Scott & Schulz, 2017)

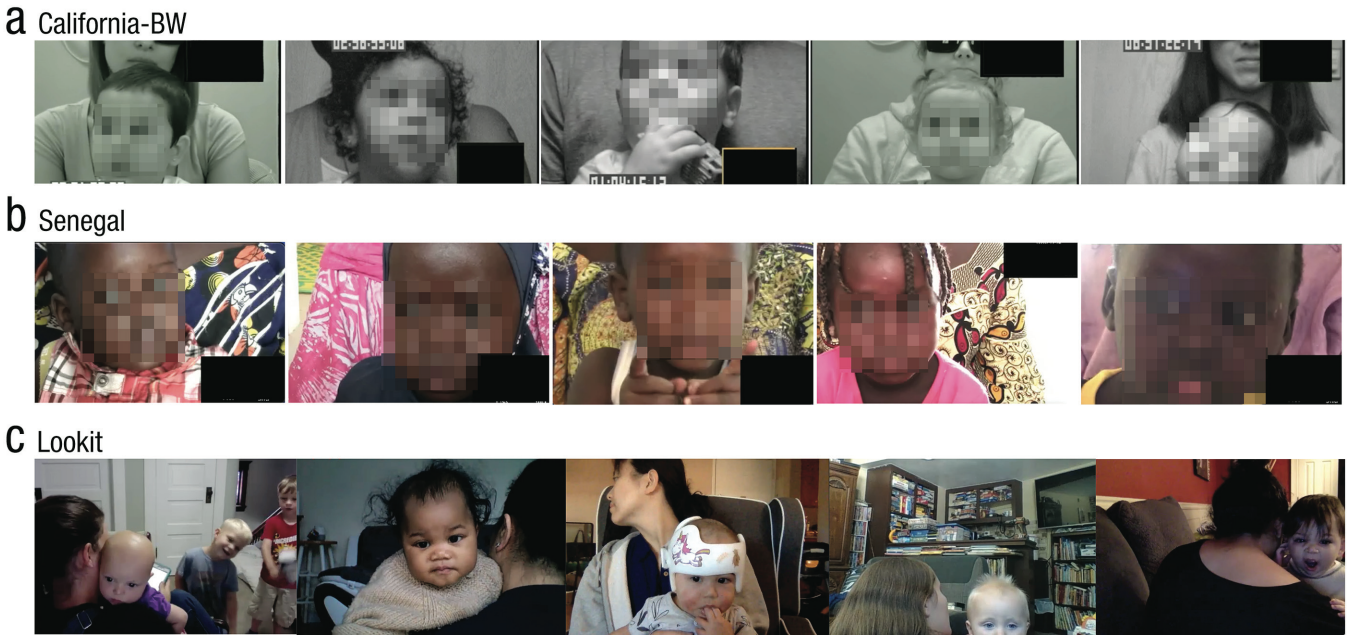

Fig. 1. Still frames from the (a) California Black and White (California-BW), (b) Senegal, and (c) Lookit data sets. Children's faces in Figs. 1a and 1b have been blurred to protect participant identity. Families featured in Fig. 1c gave explicit permission for pictures and videos to be shared for public use.

and tolerate the many sources of variability in the resulting videos.

Second, iCatcher needs to perform with high accuracy for children of varying age, race, and ethnicity. The video data set in Erel et al. (2022) included children ages 1.5 to 6 years in a majority-White sample recruited from one geographic area, leaving open possible gaps in performance for younger infants or children of different races and ethnicities. Given that looking behavior is a predominant dependent measure used for studies of infants in the first 18 months of life (e.g., Aslin, 2007, Oakes, 2012), and perhaps the easiest video-based measure to implement in large-scale unmoderated web experiments, it is vital to investigate whether iCatcher can be used to study a diverse range of infants within their first year of life. And although the field of developmental psychology tends to study White children from middle- or upper-class backgrounds (Roberts et al., 2020), online research has the potential to enable many more families to participate in science by lowering the time and energy cost for participation. In our view, iCatcher should support this goal of broadening participation and thus be held to a standard of robustness for children of varying demographics (i.e., a tool that provides human-like accuracy for White infants but not participants of other races is not a usable tool).

Third, iCatcher should deliver accurate annotations of looking behavior for studies across experimental paradigms, research questions, dependent variables, and annotation guidelines. Our goal, therefore, was to develop, train, and test a new version of iCatcher (referred to hereafter as iCatcher+) on three data sets collected in substantially different settings (online vs. in the lab vs. outside of the lab), on different topics (intuitive physics vs. language comprehension), in participants varying in age and race/ethnicity, and with different protocols for annotating looking behavior.

Present Research

In sum, iCatcher holds promise for solving the problem of automated gaze annotation from videos of infant and child participants, but its accuracy has not been tested on more diverse and challenging data sets, and its performance has not been evaluated in the terms most relevant to researchers in developmental psychology. Here, we tackle these aims and show that iCatcher+ can be used to reliably annotate infants' and young children's looking behavior at home, in the lab, and in the field; in participants of varying race, ethnicity, and age; and in videos that vary substantially in background, screen size, viewing distance, participant pose, and luminance. In anticipation of the challenges presented by the three data sets in this work, we made several technical improvements to the architecture presented by Erel et al. (2022). Then, we subjected iCatcher+ to a training and testing regime that balances key participant demographic variables such as age, race/ethnicity, and gender. The network was trained

from human annotations to classify looking behaviors as directed toward the left or right side of the stimulus display (LEFT vs. RIGHT) or away from the stimuli (AWAY) and then tested on held-out videos.

We show that the network generalizes to held-out videos from the same data set as training, with near-human trial-level performance for LEFT and RIGHT looks (with room to grow for frame-by-frame performance for the data set collected online) and lower performance for AWAY. In addition, we show that the failure modes and confidence scores produced by the model are interpretable, which allows iCatcher+ to be incorporated in more efficient machine-assisted research protocols. Finally, we show that the network generalizes not only to held-out videos from the same data set it was trained on but also to videos from a novel, fourth data set collected online using different methods and stimuli. Throughout the article, we present the model's performance in terms relevant to developmental psychologists, including frame-by-frame, trial-level, and video-level comparisons between human-to-human reliability and human-to-iCatcher+ reliability. We end by discussing the potential impact of this tool for the field of developmental psychology, giving recommendations for developmental labs hoping to adopt this tool, and previewing steps to further improve accuracy and generalizability.

Method

Nontechnical overview of approach

There are two key tasks in the current research: (a) improving the iCatcher model and designing a training regime appropriate for the video classification problem at hand and (b) testing its performance on videos it has not seen before. In this section, we define key terms and describe the iCatcher+ model and procedures for evaluating it in general, nontechnical terms. A more detailed technical description of the architecture can be found in Erel et al. (2022) and the following section titled "Model Overview."

"Model structure" consists of defining a specific task for the model to perform, breaking the problem of solving it into a series of subproblems, and then designing an architecture to solve each subproblem in turn. For this project, the (human and machine) rater's task is to label whether a participant is looking toward the left or right side on the screen or away from the screen in each frame of the video. iCatcher+ does this by detecting all the potential faces in a particular video frame, choosing the face most likely to be the participant's face, extracting the pixel values and other information from that face patch, and mapping the features from a stack of five consecutive video frames to a label for the middle frame, corresponding to the participant looking LEFT, RIGHT, or AWAY (Fig. 2). These steps may give researchers an intuition for what could drive differences between model and human performance: For example, although it is trivial for humans to find a participant's face in a frame, iCatcher+ has to be trained explicitly to correctly reject face-like patches (e.g., dolls, body parts) and faces that are not the participant's face (e.g., faces of caregivers and siblings).

"Model training" consists of tuning the network to the specific kinds of data it is learning to label, by using feedback to adjust model weights. During training, we compared the model's labels with human rater(s) trained to reliably perform this task and assume the human can generate the correct label. Feedback consists of comparing the network's guesses with the ground truth (the human rater's label) and updating its weights in an effort to increase accuracy. After model training is "model evaluation." During model evaluation, we provided iCatcher+ with new frames to label with no further feedback and compared its responses with human raters. This allowed us to test the accuracy of the model and to collect information about its failure modes (the information it has failed to learn during training).

There are several important considerations for model evaluation. First, the model needs to be evaluated on independent data to show that the model did not just memorize the correct labels for a particular set of frames. Instead, some video and annotation data are used to train the model, and other videos are "held out" of the training set so that they can be used for testing the model. Second, because we wanted to compare model and human reliability, we no longer assumed that a single human rater is 100% correct during testing. In fact, human raters do not agree 100% of the time, and the extent to which two trained raters disagree provides a data-set-specific benchmark for performance (hard data sets have lower interrater reliability; see Table 2 in Results section). Instead of comparing human and model annotations, we compared two kinds of agreement: agreement between two human raters (human–human) and between a human rater and the model (human–model). Human–human agreement is already a common measure of reliability in developmental psychology.

How different should the evaluation videos be from the training videos? As aforementioned, testing the model on the same frames it was trained on is not useful for evaluating generalization (i.e., how well iCatcher+ classifies new videos). At the other extreme, it is not reasonable, at least in early stages, to expect iCatcher+ to reliably annotate videos that are from a completely different distribution (e.g., comparing performance on videos of newborns tested in a crib with 4-year-old

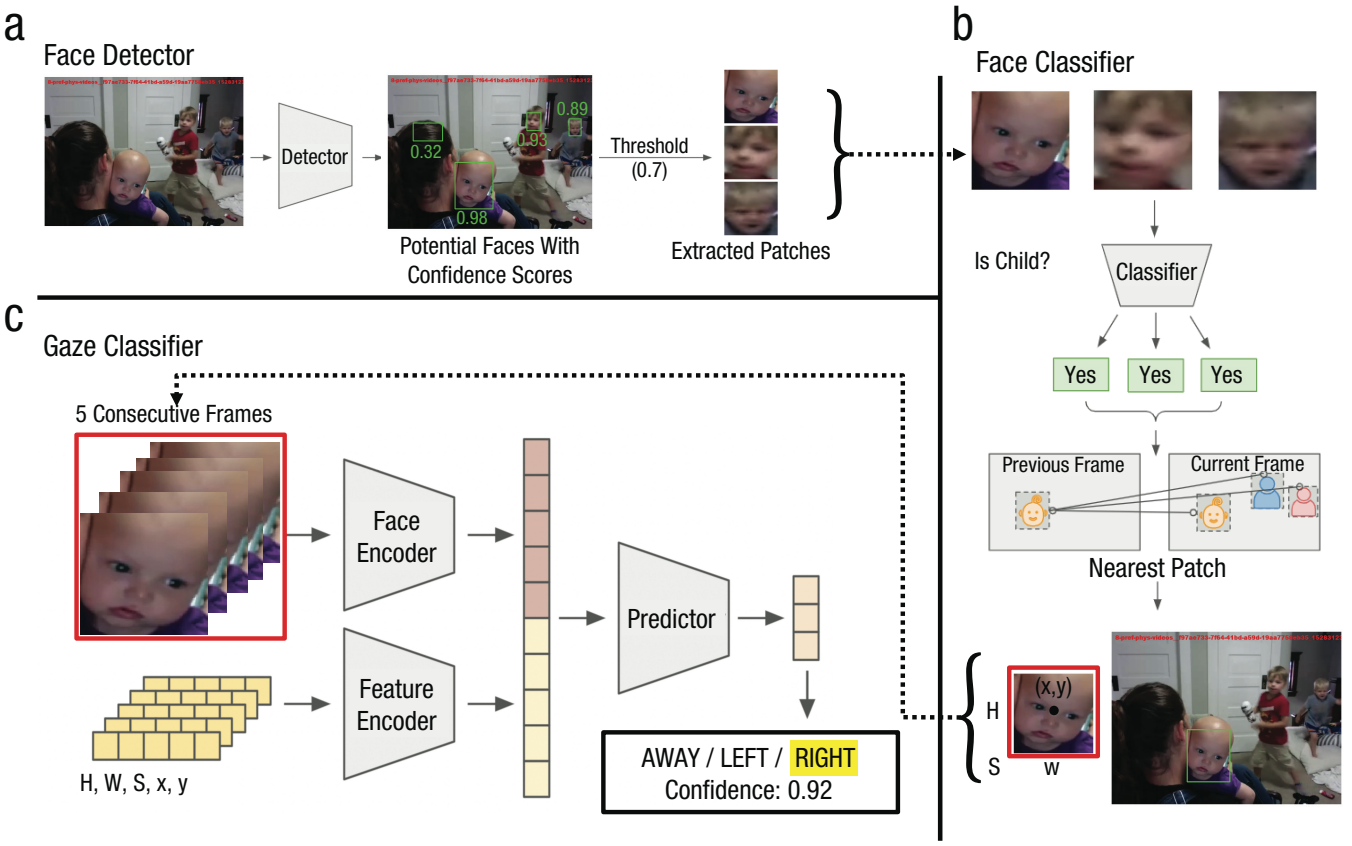

Fig. 2. Overview of iCatcher+ architecture, including the (a) face detector, (b) face classifier, and (c) gaze classifier, using frames from the Lookit data set as an example.

children tested outdoors on a playground). Thus, we began with a test of narrow generalization by evaluating the model on held-out videos within each of the three data sets. We then present one case study of far generalization in which the model trained on webcam videos collected asynchronously on the Lookit platform is tested on a separate data set of webcam videos collected via synchronous video conferencing.

Finally, the question of how well the model performs relative to a human rater can only be answered with respect to specific outcome metrics and specific measures of interrater reliability: Although the model generates a label for every frame, this is rarely the actual dependent measure of interest. In this article, we use outcome measures familiar to developmental psychologists, such as total looking time over the course of a trial or percentage looking to the right or left side of the screen. We show that trial-level human–model reliability is within the range of human–human reliability reported in studies from developmental psychology, and thus the model can be considered as reliable as a trained human annotator when it comes to trial-level measures. In the following sections, we provide a more detailed overview of the data sets, model, and training and testing regime.

Data set overview

Past research has already shown that iCatcher can be trained to reliably classify one video data set, but for the current research, we wanted to include video data sets with different and more variable video and participant characteristics. We chose three video data sets, one collected in the lab and in the field using a mobile testing trailer in the United States (California Black and White Video [California-BW]), one collected in the field using a mobile testing tent in Senegal (Senegal), and one from the Lookit online platform (Lookit), across a sample of infants and young children ages 4 months to 3.5 years. An overview of the three data sets is shown in Table 1, and example frames from each data set are shown in Figure 1. We briefly discuss the features of the entirely held-out data set, collected via Zoom, used in this article to test for far generalization in the Results section.

Table 1. Overview of Data Sets

	California Black and White Video	Senegal	Lookit
Number of children and videos	214 children, 214 videos	143 children, 143 videos	83 children, 265 videos
Research setting	University campus lab and mobile lab brought into communities	Community spaces in participating villages	Homes of participating families
Research topic	Language comprehension	Language comprehension	Intuitive physics
Participant age range	15–39 months (M = 24.46, SD = 6.46)	20–42 months (M = 30.90, SD = 6.41)	4–14 months (M = 9.03, SD = 2.33)
Participant gender	107 (50%) female, 107 (50%) male	64 (45%) female, 79 (55%) male	44 (53%) female, 39 (47%) male
Participant race/ethnicity	112 White, 42 multiracial, 42 Latine, 13 Asian, 5 Black	143 Black	62 White, 15 multiracial, 3 Latine, 3 Asian, 1 Black
Participant posture	Children sitting in caregivers' lap	Children sitting in caregivers' lap	Children mostly held over caregivers' shoulder (94% of videos with this starting position)
Video characteristics	Black and white videos, 720×480 pixels	Color videos, 640×400 pixels	Color videos, 640×480 pixels
Screen characteristics	Stimuli presented at 36 × 50 cm per picture in the university lab and at a smaller standardized ratio in the mobile lab; participants seated 3 ft away	Stimuli presented on a 17-in. laptop at a viewing distance of approximately 2 ft	Variable screen size (laptop and desktop computer screens) and viewing distance

One challenge in automating gaze classification is that every lab follows idiosyncratic annotation procedures. Because the model learns to infer exactly three classes of behavior (i.e., looking LEFT, RIGHT, and AWAY), it is important to understand how the human annotations were generated in each data set and how these annotations should be mapped to each of the three categories. The manuals used to train human annotators for the three data sets are openly available at https://osf.io/zgcb9/ (California-BW, Senegal) and https://osf.io/42hpq (Lookit). Each human rater was trained using these manuals on a set of example videos until they obtained at least 90% frame-by-frame agreement before working on the videos in this data set. Across all data sets, human raters were naive to experimental conditions and the stimuli displayed and annotated each video independently from other raters. Disagreements between raters were not resolved before training and testing with iCatcher+. Each data set also specifies frames to be included for analysis: calibration sequences and experimental trials in which stimuli were shown on the screen, and participant gaze direction is of analytical interest. We excluded from analysis all other video segments, including setup time, intertrial intervals, and pauses, which did not contain relevant looking behaviors (see Fig. 3a, white portions of timeline for Human 1 and Human 2 annotations).

Below, we provide a brief overview of each data set, including specific definitions for the three classes of looking behaviors to be learned by iCatcher+ (for details, see Tables S2 and S3 in the Supplemental Material available online).

California-BW. The California-BW data set, including 214 videos of 214 English- and Spanish-speaking children, was aggregated from 15 studies conducted in northern California. The studies measured young children's gaze behavior to study their real-time word comprehension. In the looking-while-listening (LWL) procedure (Fernald et al., 2008), children view pairs of pictures (e.g., ball and cookie) on a screen and listen as one of the pictures is named ("Where is the ball?"). Looks to the target stimulus from the onset of the key disambiguating word (e.g., "ball" in "Where is the ball?") yield high-resolution measures of speech-processing efficiency and comprehension. Studies using the LWL procedure have shown that infants' speech-processing efficiency increases dramatically over the course of the second year (Fernald et al., 1998) and that individual differences in speed of language processing are related to later verbal and nonverbal skills (Fernald & Marchman, 2012; Marchman & Fernald, 2008). All research for this data set was approved by the Stanford University Institutional Review Board, and informed consent was obtained from a parent or guardian.

Children were tested in a dark and quiet room in a developmental lab or mobile testing space (a recreational vehicle retrofitted for LWL data collection) and video

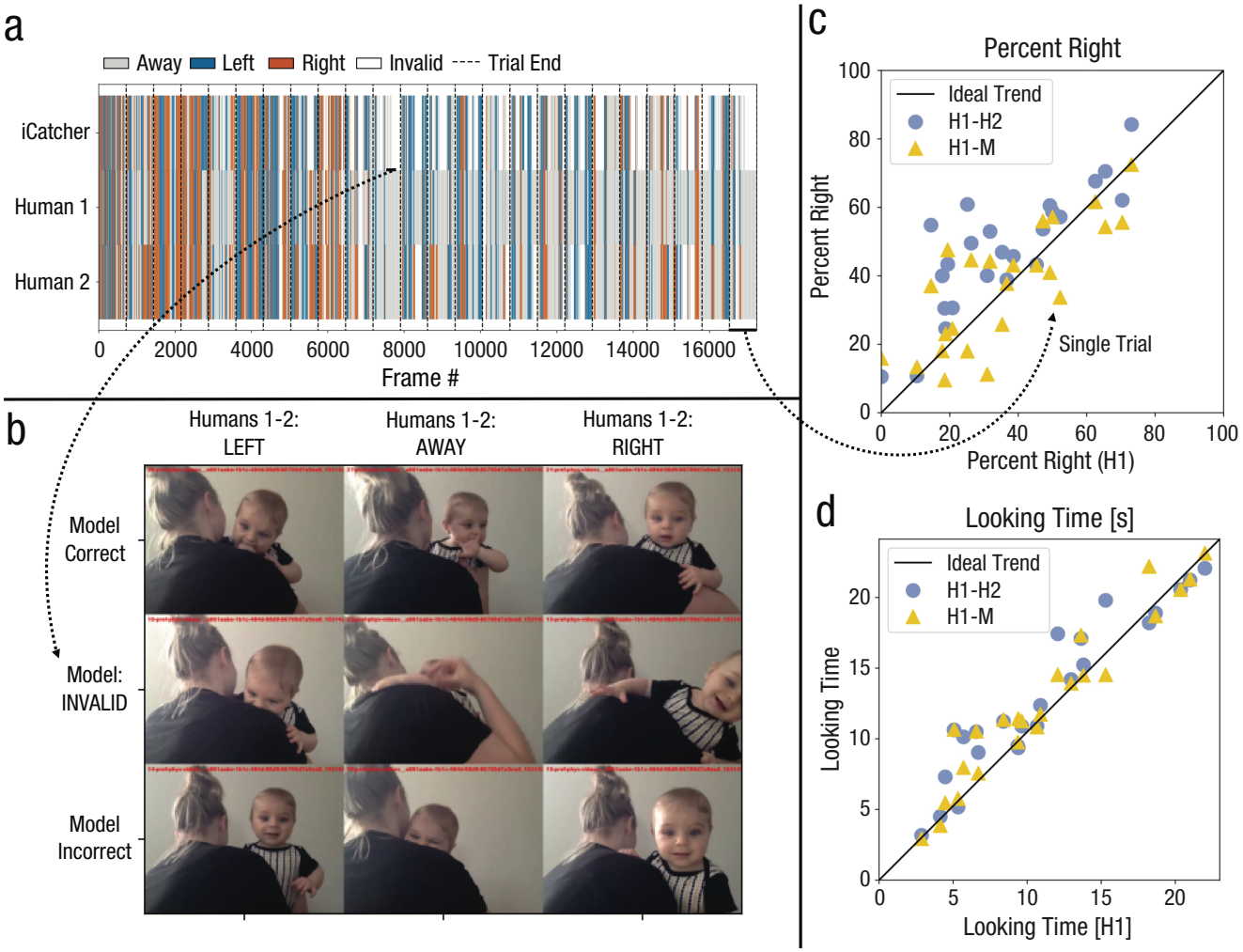

Fig. 3. Visualization of data, dependent measures, and agreement in one representative video from the Lookit data set. (a) Frame-by-frame and (c–d) trial-level agreement between human raters (Human 1 And Human 2) and iCatcher+. Figure 1a shows labels generated by iCatcher+ and by Humans 1 and 2 across all frames in the video, with trial endings marked by vertical dotted lines, and INVALID frames indicated in white. (b) Example frames in which both human raters provided a rating of LEFT (first column), AWAY (middle column), and RIGHT (last column) and iCatcher+ correctly labeled these frames (first row), provided a label of INVALID (middle row; in this case, no participant face was found, likely because of occlusion), and incorrect labels (bottom row). For plots for all videos in the test set, see https://github.com/yoterel/icatcher_plus/tree/master/plots.

recorded with a night-vision camera, yielding black-and-white videos of children's gaze patterns. Auditory stimuli were produced by a native English- or Spanish-speaking female speaker in a friendly, child-directed register. Visual stimuli were projected onto either side of a screen at a size of 36 × 50 cm per picture at the developmental lab and presented at a smaller ratio on a 55-in. LED screen at the mobile testing space. Children were seated on their caregiver's lap approximately 3 ft away from the displays. Cameras were placed in the center directly below the screen, roughly at the child's eye level. Adults wore dark glasses made opaque with dark tape over the lenses to minimize caregiver interference. Before the session, experimenters helped the caregiver and child to get seated in the testing space. During the study, the experimenter observed the child and caregiver in a control booth and could communicate with the particiIants if needed through an intercom (see Fig. 1a).

More than 25 research assistants contributed to human annotation of eye gaze across the 15 studies in this data set. Each child contributed one video session of 24 to 48 trials, and each trial lasted between 3 and 6 s. Human raters manually labeled each 33-ms frame as looking to the left or right picture (LEFT vs. RIGHT for iCatcher+) or as away or off during gaze shifts between pictures and during looks away from the screen (both mapped to the AWAY label). Trials that were excluded because of child inattention, experimenter error, or external interference were prescreened by human raters before annotation, and thus all the annotations available for training

and test come from relatively "clean" trials (for prescreening protocol, see https://osf.io/dshkr). All included trials for each video were annotated by two research assistants to assess interrater reliability. The detailed annotation scheme is shared at https://osf.io/zgcb9/ and in Table S2 in the Supplemental Material.

Senegal. The Senegal data set was drawn from a field-based longitudinal study that assessed real-time word comprehension using the LWL procedure described above. Data were collected in partnership with a local nongovernmental organization and participating Wolof-speaking villages located in a single rural region of Senegal. All research for this data set was approved by the Stanford University Institutional Review Board, and informed consent was obtained from a parent or guardian.

Children were assessed in a quiet and low-traffic indoor community space within each village, although the level of background activity varied across testing sites. To reduce visual distraction, a portable 5 × 5 ft cabana enclosed a small table and a 17-in. laptop computer presenting stimuli. The laptop keyboard was obscured with a black cover. Children were seated on their caregiver's lap at the entrance of the cabana approximately 2 ft from the laptop display. Visual and auditory stimuli were designed to be appropriate for the region (e.g., images showing local animals, objects, and food described in the Wolof language). Two portable speakers placed behind the laptop played recordings of sentences produced by a native Wolof-speaking female speaker using a register judged to be appropriate for children of this age range. Children's gaze behavior was recorded on a camcorder positioned with a tripod behind and above the center of the laptop screen. An experimenter was present to provide technical support and instructions to participating families. Caregivers wore opaque glasses to obscure visual stimuli and reduce the potential for interference (see Fig. 1b).

A Senegalese team of research assistants prescreened trials for exclusionary criteria and manually rated each 33-ms frame as LEFT, RIGHT, or AWAY following the protocol described for the California-BW data set above. Details for this prescreening and annotation scheme are available at https://osf.io/dshkr and https://osf.io/zgcb9/, respectively, and in Table S2 in the Supplemental Material. Approximately half of all included trials for each video were annotated by two research assistants to measure interrater reliability.

Lookit. The Lookit data set includes 265 videos of 83 infants, tested at home via the Lookit platform (Scott & Schulz, 2017), in a study of physical understanding. This study was designed to use the Lookit online developmental lab to conduct dense repeated sampling of infants' looking behavior. Families were invited to participate in as many as 12 sessions over 2 months. The primary measure of this study was preferential looking (left vs. right) to videos that violated a previously documented early-emerging physical expectation (e.g., unsupported objects fall, objects are solid; Baillargeon et al., 2016; Spelke et al., 1992). In each trial, two videos played simultaneously, one on the left side and one on the right side of the screen. The videos showed a single object in an event that was either physically plausible (e.g., a hand places a ball in the middle of an inclined ramp, releases it, and the ball accelerates down the ramp) or physically implausible (e.g., upon release, the ball accelerates up the ramp). The study was approved by the Massachusetts Institute of Technology Institutional Review Board, and informed consent was obtained from a parent or guardian before participation.

In each video from the Lookit data set, participants saw up to twenty-four 20-s trials while a webcam recorded their looking behaviors. The video recording of each session was a concatenation of separate recordings from each portion of the experiment (e.g., parental consent, setup, each trial). No experimenter was present to provide synchronous guidance to caregivers. Instead, caregivers were provided with detailed instructions for how to set up the study, such as ensuring infants' faces were captured in the webcam feed and illuminated from the front with minimal backlighting. Caregivers were instructed to face away from the screen and to hold infants over their shoulders (this was the starting posture for 94% of videos), although infants did, in rare cases, sit on their caregiver's lap or in chairs by themselves. This resulted in videos that varied significantly in infant position, viewing distance and video illumination, resolution, and background (see Fig. 1c).

Twenty-four trained research assistants contributed to the annotation of this data set. Human raters manually labeled each frame (at 33-ms intervals) as directed toward the left or right of the screen (LEFT, RIGHT) or away from the screen (AWAY). Each video was annotated by one research assistant, and a random subset of videos (23%) was annotated by a second rater to assess interrater reliability. The detailed annotation scheme is available at https://osf.io/pq6ng/ and in Table S3 in the Supplemental Material.

Differences between data sets. These three data sets were selected because they differ from each other and from the original data set used in Erel et al. (2022) in important ways, including research setting (lab vs. field site vs. online), study topic (language development vs. intuitive physics), age range (infants 1 year old or younger vs. 2- and 3-year-olds), and mode of data collection (experimenter present vs. absent). In particular, the between-video variability in the Lookit data set is substantially higher than in the

California-BW and Senegal data sets in terms of screen size, backdrop, camera position, child posture and movement, viewing distance, and lighting—features that were standardized in the other two data sets. Whereas videos in the California-BW and Senegal data sets were collected at a constant frame rate (30 frames per second), all videos in the Lookit data set were collected at variable frame rates (because of the usage of webcams and other devices that do not keep a constant frame rate) but resampled to 30 frames per second for consistency before being annotated and passed to the iCatcher+ model. Because these three data sets differ from each other in many ways, it is hard to pinpoint the cause(s) of differences in classification accuracy in the absence of more closely matched data sets, ideally with random assignment (though see the Results section, wherein we explore the factors that are associated with better performance).

A last key difference between the data sets is the human-annotation protocols (for a detailed description of these annotation schemes, see Tables S2 and S3 in the Supplemental Material). All data sets define preferential looking between the left versus right side of the screen to be the primary dependent measure, but there are also subtle and important differences in the annotation schemes across data sets. First, before gaze annotation, human raters prescreened the California-BW and Senegal data sets and excluded trials with excessive participant fussiness, distraction, or caregiver interference or insufficient infant attention. The Lookit data set was not prescreened, so the annotations include trials in which the infant was fussing or inattentive. Second, the data sets differ in the treatment of looking behavior that was not directed to either the left or right stimulus but nevertheless was still directed toward the screen. Because the measures of interest in the California-BW and Senegal data sets were the reaction time of gaze shifts from a distracter to target image (e.g., gaze shift from image of cookie to ball upon hearing "Where's the ball?") and the overall the time spent looking at each of the two specific image locations, these transitional looks were annotated by humans as "off" (i.e., on the screen but off-stimulus) and mapped to the iCatcher+ label of AWAY. In contrast, all looks toward the screen in Lookit were annotated by humans as either LEFT or RIGHT (including transitional looks between stimuli). Thus, although we take looking duration as the sum of the time within a trial that a participant looks LEFT or RIGHT in all data sets, in California-BW and Senegal, this includes only frames in which the participant was looking at one of the two images. Third, in the California-BW and Senegal data sets, human raters only annotated gaze fixations that lasted at least three frames (100 ms) and only annotated gazes as off target or away from the screen that lasted at least six frames (200 ms). No such criteria were implemented in the Lookit data set; thus, gaze shifts could be more frequent or brief.

Model overview

In this section, we provide a general overview of iCatcher+ (for details, see Erel et al., 2022, and the Supplemental Material). As shown in Figure 2, iCatcher+'s model architecture consists of three major components, a face detector, a face classifier, and a gaze classifier, all operating on five consecutive frames at a time, hereafter referred to as a "data point." The goal of the system is to predict the category of gaze (LEFT, RIGHT, AWAY) for the middle frame within this moving window of five consecutive frames. During training, all data points were prepared during preprocessing, and during evaluation, data points were created on the fly, enabling the potential for annotation to occur in real time while the experiment is running. The face detector (Fig. 2a) extracts potential portions of the image that plausibly contain the participant's face. Candidate patches are then fed into the face classifier (Fig. 2b), which determines whether the patch belongs to an infant or adult, and if multiple candidate faces are found, which face is most likely to belong to the participant. The five selected patches from each data point, together with their size and x-y position in the frame, are then fed to iCatcher+'s gaze classifier (Fig. 2c). This component estimates the discrete gaze direction for the middle frame.

Face detector. Just as in Erel et al. (2022), we used the face detector provided by OpenCV (Bradski, 2000). This off-the-shelf detector was not trained by us or tuned toward extracting infant faces. In addition to returning potential face patches, the face detector outputs a confidence score between 0 and 1 for each potential face, which we used to filter out the patches using a threshold of 0.7 in all our experiments. The output from this component is a list of the upper-left and bottom-right coordinates of the pixels of each candidate patch.

Face classifier. Because the candidate patches from the face detector may contain adults, body parts, and even objects, we passed the patches through a face classifier tasked with selecting the patch most likely to contain the participant's face. To this end, a separate neural network was trained to distinguish between patches of infants and noninfants. The full architecture and training procedure of the face classifier are described in the Supplemental Material. Furthermore, we added an additional constraint to protect against selection of face patches from different people across consecutive frames. To do this, we first filtered the candidate patches to only those likely to contain infant faces. If more than one candidate patch remained,

we chose the closest patch relative to the selected patch from the previous frame (Fig. 2b). Hence it is possible, in principle, for a wrong face to be selected if the participant moves out of frame and another child's face is detected. During training, if the face classifier could not return a face patch for one or more of the five frames in the data point, we disregarded it (see Fig. S1 in the Supplemental Material). During testing, if the face classifier did not return a face, we used a placeholder of all black pixels instead, and the data point was still considered valid as long as the middle frame was judged to contain the participant's face. This behavior ensures that the data set is "clean" for training yet quite robust to missing information during evaluation (e.g., even if only one frame in a five-frame sequence contains the infant's face, we can still classify that frame).

Gaze classifier. Given a data point consisting of five consecutive face patches and their spatial information (height, width, size, and center coordinates), the gaze classifier is tasked with predicting the gaze direction of the middle frame. The direction is encoded by three discrete classes: AWAY, LEFT, and RIGHT. During training, we passed all data points from the data set through a RandAugment (Cubuk et al., 2020) block, which performs various random image-level augmentations. In randomly selected frames, we also horizontally flipped the five image patches, their respective spatial information (the Center of Patch × Coordinate), and their label (i.e., LEFT becomes RIGHT, RIGHT becomes LEFT, and AWAY is kept the same). These augmentations were not activated during evaluation. The gaze classifier itself is a neural network consisting of a feature extractor that is a pretrained ResNet18 (He et al., 2016) and a classifier that consists of three fully connected layers. Cross-entropy loss was used for optimization. The full architecture and training procedure are described in the Supplemental Material.

Classification output. Given a video, iCatcher+ returns frame-by-frame labels (LEFT, RIGHT, AWAY) for all frames within that video in which a participant's face was identified and a confidence score for each class that sums up to 1 (e.g., 0.1 for LEFT, 0.8 for RIGHT, and 0.1 for AWAY). For frames in which no face was identified, iCatcher+ returns a label of INVALID, which can be broken down into the subcategories of NOFACE (if no faces were detected at all in that frame) and NOBABYFACE (if faces were detected by the face detector but no participant faces were found by the face classifier). For illustrations of this output overlaid on video data, see https://osf.io/frmgx/.

Data set splitting for training, validation, and test

In the previous sections, we described iCatcher+'s architecture and the three video data sets that are the focus of the current work. In this section, we describe a procedure for splitting the data to evaluate model performance. A "split" in this context is an assignment of each video in the data set to one of the following subsets: "training," "validation," and "test." We trained iCatcher+ on the training set, then assessed the quality of the training procedure using the validation set. The performance on the validation set provides a rough estimate for the performance on the test set during training but is not included in any training or in the final results. The test set consists of unseen videos, and performance of the trained model on this test set provides a proxy for model performance on new videos that are similar to the training distribution. The test set was not used to improve iCatcher+ either directly (optimization) or indirectly (hyper-parameter tuning), thus ensuring that performance on the test set is driven only by the model's ability to generalize from training.

We used stratified random sampling to assign videos to the training, validation, and test sets. First, we divided individual infant participants into mutually exclusive strata defined by all possible combinations of key demographic variables. For example, one stratum in the Lookit data set was defined by "4-6-months old, White, females." Next, within each stratum, we assigned a fixed proportion of infants into a training set (approximately 70% for California-BW and Senegal; 80% for Lookit[1]) and the remaining infants to a test set. We further sampled 10% of the infants within the training set for validation (for final counts, see Table 2). This procedure ensured that the multivariate demographic distribution in each data set reflects that of the overall data set, and that the test set included infants that did not appear in the training set. The California-BW data set was split according to age (in 4-month bins), gender (male or female), race/ethnicity (White: $n = 112$, 52%; other: $n = 102$, 48%), and by preterm birth (<33 weeks' gestation: $n = 72$ or 34%), which is a plausible predictor of delays in early language processing. The Senegal data set, consisting of Black Senegalese children, was split according to age in 4-month bins and gender (male or female). The Lookit data set was split according to age (in 3-month bins because of narrower range, using earliest age of participation for infants who contributed data to multiple videos), gender (male or female), and race/ethnicity (White: $n = 62$, 75%; other: $n = 21$, 25%).

We preprocessed each data set to maximize information for model training and evaluation. For model training, we excluded frames for which human raters disagreed (for breakdown of included and excluded frames for training, see Fig. S1 in the Supplemental Material). Although all included trials in all videos in the California-BW data set were annotated by two human raters, only about half the trials in all videos from the Senegal data set and a random subset ($n = 16$) of videos

Table 2. Summary of Main Results

Data set	*N* videos in training / validation / test (*N* trials)	% Invalid frames flagged by model [95% CI]	Comparison	% Agreement	Cohen's κ	ICC looking time	ICC percentage right
California Black and White Video (iCatcher+)	135 / 15 / 64 (1,861 trials in test set)	15.01% [11.37%, 19.14%]	H1 - H2	97.79% [97.11%, 98.32%]	0.96 [0.95, 0.97]	0.94 [0.91, 0.97]	0.99 [0.99, 1.00]
			H1 - M	95.11% [93.56%, 96.33%]	0.91 [0.89, 0.93]	0.90 [0.86, 0.94]	0.96 [0.93, 0.98]
Senegal (iCatcher+)	89 / 9 / 45 (576 trials in test set)	17.19% [13.42%, 21.54%]	H1 - H2	98.05% [97.56%, 98.51%]	0.97 [0.96, 0.97]	0.94 [0.92, 0.97]	0.99 [0.98, 1.00]
			H1 - M	90.91% [87.25%, 93.61%]	0.85 [0.80, 0.89]	0.89 [0.83, 0.94]	0.95 [0.90, 0.98]
Lookit (iCatcher+)	148 / 8 / 45 (1,026 trials in test set)	20.66% [16.59%, 25.40%]	H1 - H2	90.99% [89.57%, 92.31%]	0.85 [0.83, 0.87]	0.95 [0.93, 0.97]	0.89 [0.85, 0.92]
			H1 - M	85.23% [83.46%, 86.97%]	0.75 [0.72, 0.78]	0.95 [0.93, 0.97]	0.81 [0.73, 0.88]
Lookit (original iCatcher)	148 / 8 / 45 (1,026 trials in test set)	14.89% [11.64%, 18.69%]	H1 - H2	90.96% [89.57%, 92.31%]	0.85 [0.83, 0.87]	0.95 [0.93, 0.96]	0.89 [0.85, 0.92]
			H1 - M	73.68% [70.21%, 77.12%]	0.57 [0.52, 0.62]	0.85 [0.81, 0.88]	0.63 [0.53, 0.72]
Zoom (iCatcher+ trained on Lookit)	0 / 0 / 63 (712 trials in test set)	25.40% [20.88%, 29.62%]	H1 - M	85.87% [84.31%, 87.31%]	0.46 [0.41, 0.51]	0.97 [0.97, 0.98]	—

Note: This table shows information about the number of videos in the training/validation/test split, the number of trials in the test set that both humans (H1 and H2) and the model (M) annotated, percentage of frames that humans rated (i.e., during trials in which looking behavior was of analytic interest) but was flagged by the model as "INVALID," and human–human (H1-H2) versus human–model (H1-M) agreement at the level of frames (percentage agreement, Cohen's κ) and trials (intraclass correlation coefficient [ICC] over looking times and percentage looking to the right), averaged over videos within each data set. For the four agreement metrics, we present the mean and 95% confidence intervals (CIs) computed via bootstrapping over 1,000 iterations. Cohen's κ is affected by (a) the number of categories and (b) the distribution of observations over those categories, so comparisons across data sets on this measure should not be interpreted.

from the Lookit training data set had a second human annotator. For model evaluation on held-out videos, we designated the "primary" human rater as "Human 1" to compute human–model comparison metrics for California-BW and Senegal. Because there was no designated primary rater for the Lookit data set, when two human annotations were available, we randomly selected one of them to be Human 1.

Overview of dependent measures and measures of reliability

All videos included up to 48 (California-BW), 44 (Senegal), and 24 (Lookit) trials of data collected in a single experimental session. We used frame-by-frame annotations from human raters and iCatcher+ (Fig. 3a) to generate trial-level dependent measures most relevant to developmental researchers: preferential looking (PR; proportion looking toward the right side of the screen relative to total looking on the screen; Fig. 3c) and looking time (LT; total time looking toward the screen for Lookit or toward one of the two images for California-BW; Fig. 3d). Then we compared human–human and human–model agreement using metrics familiar to developmental researchers: percentage agreement and Cohen's κ (range = 0–1) over frames and intraclass correlation coefficient (ICC; range = 0–1) over trials. For precise definitions of these metrics, see Table S5 in the Supplemental Material. For all of our results, metrics are always presented with respect to a time interval in which they were explicitly averaged over (e.g., frames, trials, videos). Frames that were not annotated by one of the annotators (model or human) were not considered for these comparisons.

Open science practices

The work presented in this article was not formally preregistered. However, all training and technical work on the network was done without access to results from the test set, which was untouched until training and validation were complete. Throughout this project, when we noticed blatant errors in human raters' annotations or other missing information that prevented the model from parsing the annotations (e.g., no trial time stamps,

mismatches between file names of annotations and videos), we corrected them.[2] The dependent measures and reliability metrics were chosen in advance of the results, and we did not explore any other measures or metrics, except for coefficient of individual agreement (Parker et al., 2020), which we dropped because we could not find an implemented function from published work. Our confirmatory results are the comparisons between human–human and human–model agreement on the four agreement metrics (percentage agreement, Cohen's κ, ICC for looking duration and preferential looking). All data and code required to reproduce the main results and figures of the article and plots specific to each video in the test set are available at https://zenodo.org/record/7232828, which is a snapshot of the repository found at https://github.com/yoterel/icatcher_plus. The raw video files for the Lookit data set are available at https://osf.io/yfrkw/ (public data sharing) and https://osf.io/r7czb/ (scientific data sharing); all videos in this set are shared with consent from a legal guardian. The raw video files for the California-BW and Senegal data sets are not publicly available given restrictions to protect participant privacy. The data management plan and annotation files for all data sets are available at https://osf.io/ujteb/.

Results

Table 2 and Figure 4 summarize the main findings. For examples of representative good and poor performance, see Video S1 in the Supplemental Material.

Comparing human–human and human–model agreement

Below we compare human–human and human–model agreement. Note that these comparisons are computed only over frames that both humans and the model treated as valid (i.e., for the model, an infant face was detected; for humans, the infant was not distracted, other adults and children were not interfering, etc.). We report more information about invalid frames in the Evaluating Failure Modes section. See Figure 5 for a scatterplot comparing human-human agreement to human–model agreement for all three data sets.

California-BW. iCatcher+ achieved a near-human mean frame-by-frame agreement of 95.11%, bootstrapped 95% confidence interval (CI) = [93.56%, 96.33%], over all videos in the test set (vs. human–human agreement of 97.78%, 95% CI = [97.11%, 98.32%]). For trial-level metrics, iCatcher+

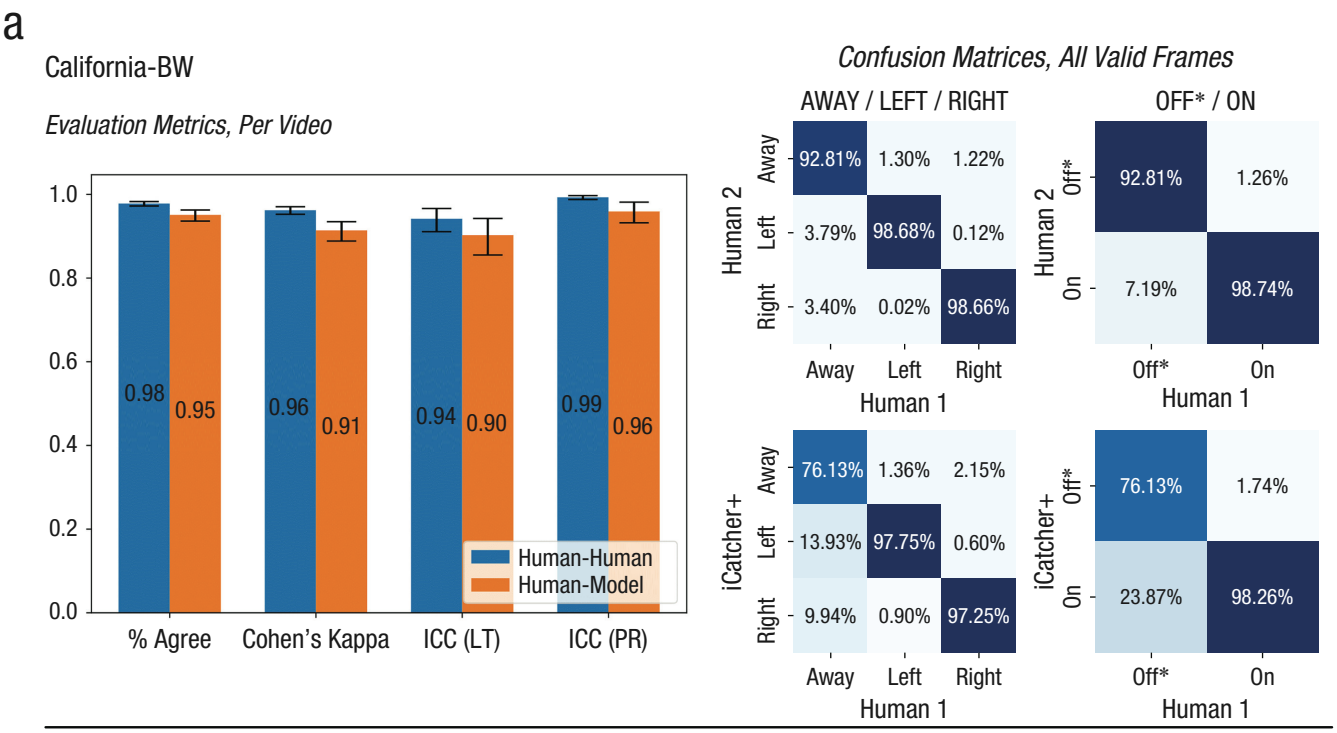

Fig. 4. *(continued on next page)*

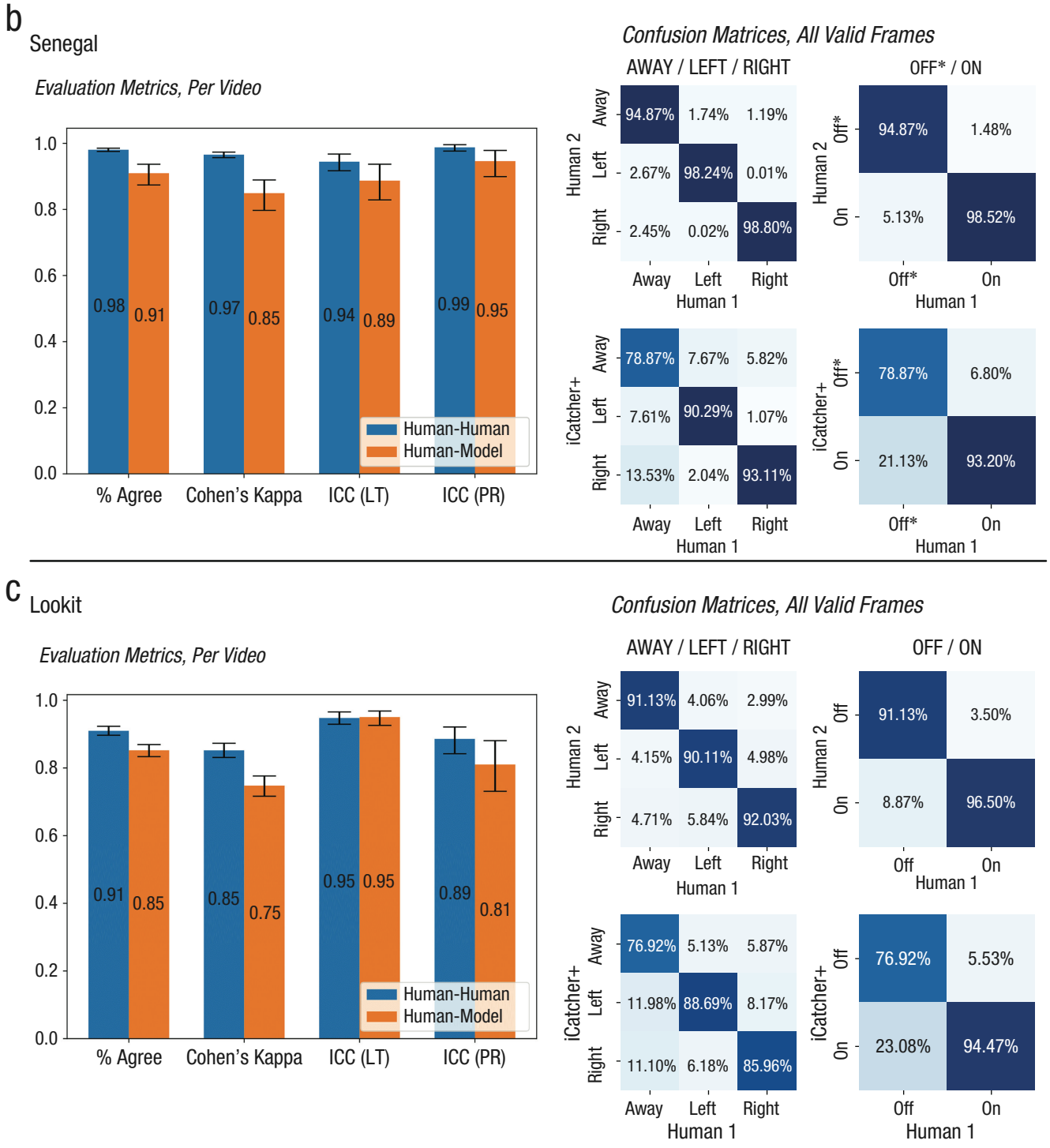

Fig. 4. Main results over all videos in the test set of (a) California Black and White Video (California-BW), (b) Senegal, and (c) Lookit, calculated over all mutually valid frames (i.e., frames for which both humans and the model provided annotations). (Left) Bar plots showing mean values on evaluation metrics (percentage agreement, Cohen's κ, and intraclass correlation coefficient [ICC] over looking time [LT] and percentage right [PR]) and bootstrapped 95% confidence intervals. Confusion matrices on the right show the proportion of frames in the test set that Human 2 agreed with Human 1 (first row) and that iCatcher+ agreed with Human 1 (second row) on classifying looks as AWAY versus LEFT versus RIGHT (left column) or OFF versus ON (right column). Note that for California-BW (a) and Senegal (b), OFF means looking at neither image (marked with *), which includes both looking away from the screen and looking in between the stimuli, versus in Lookit (c), where OFF exclusively means looking away from the screen.

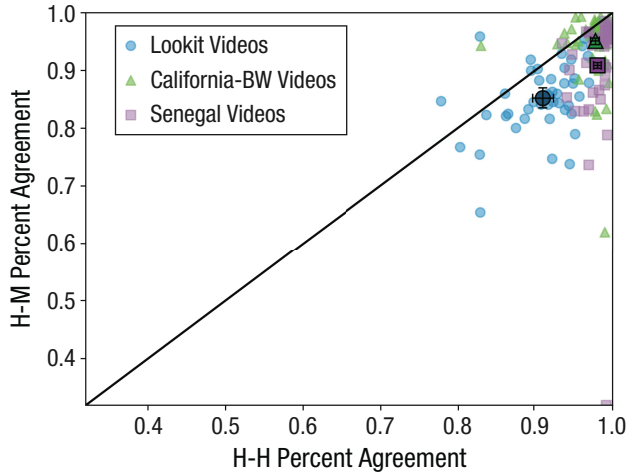

Fig. 5. Comparing human–human agreement (horizontal axis) and human–model agreement (vertical axis) across the test portions of the three data sets. Each point represents the frame-by-frame percentage agreement over all mutually valid frames (i.e., frames that both the model and the human rater annotated) per video, and the larger points and error bars indicate the mean percentage agreement, and the bootstrapped 95% confidence intervals (CIs) over all videos in the respective test sets.

achieved near-human performance, with best performance for classifying looks as LEFT or RIGHT (average ICC for percentage looking = 0.96, 95% CI = [0.93, 0.98] vs. between humans 0.99, 95% CI = [0.99, 1.00]) and lower, although still excellent, performance for classifying looks as ON (i.e., toward either image) or OFF (i.e., toward neither image; average ICC = 0.90, 95% CI = [0.86, 0.94] vs. 0.94, 95% CI = [0.91, 0.97] between humans).

Senegal. The Senegal data set, relative to the California-BW data set, included more variability in setting and lighting; the viewing distance was also shorter (2 ft vs. 3 ft). iCatcher+ achieved an average of 91% frame-by-frame agreement over videos in the test set (90.91%, 95% CI = [87.25%, 93.61%]) versus human–human agreement of 98.05% (95% CI = [97.56%, 98.51%]). At the level of trials, iCatcher+ achieved near-human performance on LEFT or RIGHT classification (ICC = 0.95, 95% CI = [0.90, 0.98] vs. between humans, ICC = 0.99, 95% CI = [0.98, 1.00]) and lower but still excellent performance for ON or OFF classification (ICC = 0.89, 95% CI = [0.83, 0.94] vs. between humans, ICC = 0.94, 95% CI = [0.92, 0.97]). For both California-BW and Senegal, iCatcher+ was trained to classify looks in between the two images as AWAY, which the model successfully generalized to new videos.

Lookit. The Lookit data set, relative to California-BW and Senegal, included more sources of variability (lighting, resolution, camera angle and position, screen size, distance and position of the participant, background, trials during which the infant was fussy or distracted). Both human–human and human–model frame-by-frame agreement were lower for Lookit than for the two other data sets (see Table 2). iCatcher+ achieved a mean frame-by-frame agreement of 85.23% (95% CI = [83.46%, 86.97%]) over all videos (vs. human–human agreement of 90.99%, 95% CI = [89.57%, 92.31%]) and a trial-level ICC of 0.95 (95% CI = [0.93, 0.97]) for LT (vs. human–human ICC of 0.95, 95% CI = [0.93, 0.97]) and 0.81 (95% CI = [0.73, 0.88]) for percentage looking to the right (vs. human–human ICC of 0.89, 95% CI = [0.85, 0.92]).

In contrast to the California-BW and Senegal data sets, in the Lookit data set, both humans and the model were more accurate at classifying looks as ON versus OFF than LEFT versus RIGHT (human–human average ICCs for looking duration: 0.95, 95% CI = [0.93, 0.97] vs. for direction: 0.89, 95% CI = [0.85, 0.92], $t(44) = 3.45$, $p = .0012$, two-tailed, paired t test; human–model: 0.95, 95% CI = [0.93, 0.97] vs. 0.81, 95% CI = [0.73, 0.88], $t(44) = 3.35$, $p = .0017$, two-tailed, paired t test). We speculate that it was easier for humans and the model to tell whether the infant was looking at the screen than whether the infant was looking left or right, which involved classifying ambiguous frames when the infants were transitioning from looking at one side of the screen to the other.

When viewing distance and angle, camera and video resolution, and testing environment were held relatively constant in the California-BW and Senegal data sets, iCatcher+ showed excellent performance. Although performance was somewhat worse for the Lookit data set, the fact that it still approached human-level performance is striking because of the vast variability in viewing distance and angle, screen size, camera resolution, and participant positioning. We note that this is the intended goal of training: to introduce variability that the model is likely to see later on and to fine-tune the model to be robust to these features.

Unlike California-BW and Senegal, which consisted of videos from entire experimental sessions, the Lookit data set contained videos from individual trials that were concatenated during postprocessing. Thus, when the model averages across a moving window of five frames for classification, this moving window within trials contains frames continuous in time, but the same moving window across trials contains frames that skip across intertrial intervals and thus can introduce discontinuity in participant pose, gaze behavior, background, and camera angle. To explore whether this lowered performance, we calculated the average agreement for between-trials versus within-trials intervals per video. Accuracy was indeed lower for between-trials intervals: When the moving window scrolled over a trial boundary, frame-by-frame agreement dropped by an average of 5%, $t(44) = 4.88$, $p < .001$, two-tailed, paired t test (see Fig. S10 in the Supplemental Material).

Comparing iCatcher+ with original iCatcher. We also compared the performance of iCatcher+ with the original version of the model published in Erel et al. (2022) on the Lookit data set. We focused on the online Lookit data set for this comparison because the original iCatcher model has already been tested on a lab data set in Erel et al. We found that on every metric, iCatcher+ outperformed the original model (see Table 2). This suggests that the technical improvements of the model—including (a) features such as participants' face position and distance relative to the camera and random augmentations of frames during training and (b) an improved face classifier—are important for robust and accurate classification in videos collected online.

Predicting performance given participant, video, and model outputs

In the previous section, we reported that iCatcher+ achieved good performance for two data sets collected by a researcher in a lab or in the field in older infants and toddlers (California-BW and Senegal) and a data set collected without a researcher present at all in a home setting in younger infants (Lookit). To assess the viability of this model as a tool that developmental psychologists can use, we care not only about average performance across the entire test set but also about performance over variability in participant characteristics such as age, gender, race/ethnicity, skin tone, and eye color; participant behaviors such as head movement and position; and video-level features such as luminance and pixel density.

Figures S6 through S8 in the Supplemental Material show all of these features plotted against human–model agreement, and Figure S9 in the Supplemental Material shows all of these features plotted against human–human agreement. Descriptively, average performance was reasonably robust (> 80% agreement) when looking at variability for each feature. Although there is still room to grow for frame-by-frame classification accuracy for the Lookit data set (although note human-level performance for both preferential and duration looking in Table 2), these findings suggest that the gap in performance between the model and human raters is not driven only by lower accuracy on videos with participants of a certain demographic or videos with a certain set of image-level features.

For every annotated frame, iCatcher+ generates a label and a confidence score. One open question is whether this learned confidence metric actually corresponds to model accuracy. If this confidence score can be used as a way of separating easy versus difficult frames, then iCatcher+ should be less confident for videos that it classified incorrectly. This was true for all three data sets (see Fig. 6): iCatcher+ provided higher confidence scores for frames that it ultimately labeled correctly than frames that it labeled incorrectly; Lookit: $t(1698.21) = 56.32$, $p < .001$; California-BW: $t(2321.22) = 63.26$, $p < .001$; Senegal: $t(1325.50) = 48.02$, $p \leq .001$ (all two-tailed, Welch's t test, p values generated via permutation).

We ran lasso regressions (Tibshirani, 1996) to explore which predictors were most important for predicting model performance for each data set and across data sets. For this analysis, we chose to study model performance at its finest possible level of granularity (frame-by-frame percentage agreement in every trial) to maximize sensitivity. We found that participant and video features, such as face position and density, and human–human agreement were selected as important predictors. Nevertheless, a simple statistical model with iCatcher+'s reported confidence as the sole predictor (along with a random intercept for subject) already explained 62.9% (conditional R^2) of the variance in human–model agreement compared with 67.0% of explained variance in the "best" model, suggesting that other predictors explain only a small portion (≈4%) of the unique variance in model performance. For details, see the Supplemental Material. On balance, these results show that we achieved near-human performance on the dependent measures that matter most to developmental psychologists (LT, percentage looking to the left vs. right, that aggregates over frames within a trial), and that confidence scores can be used to identify trials that are likely to contain inaccurate annotations. In future work, these scores could guide decisions about whether and when automated gaze coding should be supplemented with human annotation. For example, a future tool can allow researchers to specify a confidence threshold—all trials or videos that fall under this threshold can then be flagged for closer inspection by humans.

Evaluating failure modes

In the two above sections, we reported good model performance in all data sets across variability in video and participant characteristics. We also showed that confidence scores from the model are interpretable and strongly predict accuracy. In this section, we dig deeper into the model's failure modes in a qualitative way. For each data set, we inspected the human and model annotations for 40 videos with low frame-by-frame accuracy (lowest 15 for California-BW and Lookit, lowest 10 for Senegal). For each video, we identified frames in which the iCatcher+ classification of looking behavior differed from one or both human raters. We then inspected the video segments corresponding with these frames and summarized what was happening in the video during these segments and what may have caused the disagreements between iCatcher+ and the human raters (for the

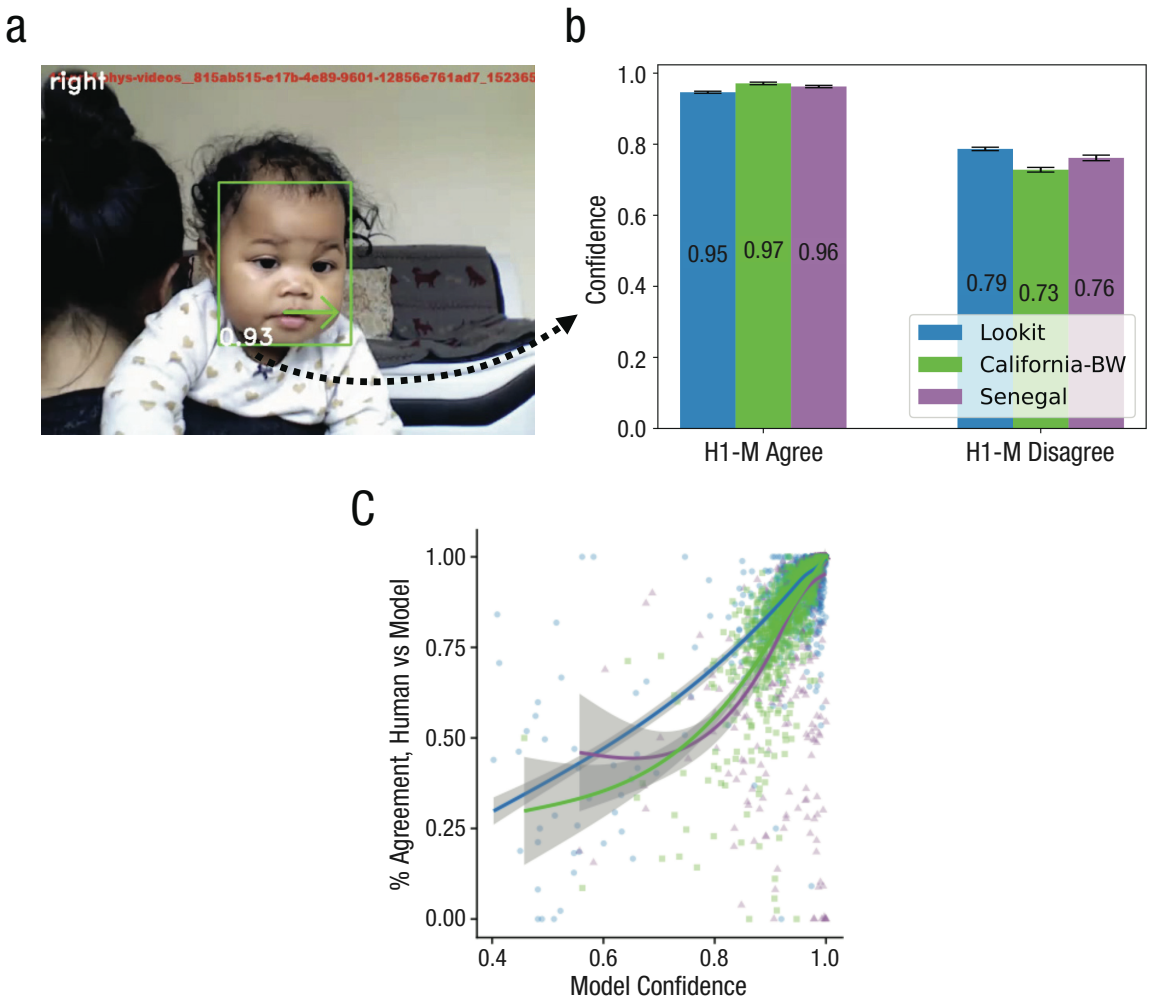

Fig. 6. (a) Visualization of the inferred face (green bounding box), label (upper left corner, also indicated by arrow), and confidence score (inside bounding box) overlaid on a video frame. (b) Confidence scores, computed per trial, for frames that the model ultimately labeled correctly (i.e., agreeing with humans, H1-M Agree) versus incorrectly (H1-M Disagree) across all data sets. Error bars indicate bootstrapped 95% confidence intervals. (c) Relating model confidence to human–model agreement on every trial. Ribbon and 95% confidence interval computed using method = 'gam' and formula 'y ~ s(x, bs = "cs")'.

raw, tabulated, qualitative data, see https://osf.io/zquys). Overall, we found two general failure cases (no faces detected, faces detected but labeling was incorrect), which we describe below. We also observed cases in which iCatcher+ classifications agreed with annotations from Human 2 more so than from Human 1 when the two human coders did not agree (agreement and other metrics were always calculated between Human 1 and the model).

No face found ("INVALID"). The first failure case is when iCatcher+ failed to detect any face in the frame and thus could not even get started with annotation (average proportion of ≈15%–20% of frames in each data set; see Table 2). Through visual inspection, a majority of invalid frames were frames in which the participant was partially or completely turned away from the screen or the participant's face was partially or completely occluded (e.g., due to participants rubbing their eyes, putting their hands in their mouths, lowering their face into a caregiver's shoulder, or cases where participants were positioned partially or fully outside of the camera's view, and moved so close to camera that parts of their faces were off screen; for examples from one Lookit video, see Fig. 3b). Because iCatcher+ needs to detect a face first in order to return a label (AWAY/LEFT/RIGHT) for that face, the model returns a label of INVALID for these frames; in fact, in many of these cases, the participant was not looking toward the screen. A smaller portion of invalid frames was from videos with dark lighting or shadows that made infants' eye movements difficult to track. There were also a few cases with poor video quality (e.g., the video was grainy or part of the video frame was blurry; see Video S1 in the Supplemental Material).

To test whether these cases, in part, drive underperformance of the model, we compared frame-by-frame agreement over all valid frames in the test set when we replaced all frames labeled by iCatcher+ as INVALID (i.e., NOFACE or NOBABYFACE) with the label AWAY. Across all data sets, we found that replacing INVALID frames with AWAY increased ON versus OFF performance (for agreement on AWAY, from 76.13% to 83.04% for California-BW, 78.87% to 82.41% for Senegal, and 76.92% to 88.82% for Lookit), at the expense of a 5% to 10% hit in performance on classifying LEFT versus RIGHT (see Figs. S3–S5 in the Supplemental Material). Thus, for all three data sets, iCatcher+ can either achieve near-human frame-by-frame performance on LEFT versus RIGHT classification or near-human performance on ON versus OFF classification but currently cannot accomplish both simultaneously.

Face found but incorrect label. The remaining failure cases are those in which the face selector and classifier identified a face but outputted an incorrect label. We found that these failure cases could be roughly categorized as (a) edge looking behavior and (b) movement. First, there were instances in which the infant looked very far to the left or right such that the model classified these looks as AWAY, whereas human raters classified these looks as LEFT versus RIGHT, or in Lookit, cases in which the infant was looking toward the center of the screen, in between the stimuli, and the model and humans disagreed about whether that look corresponded to looking in the right versus left side of the screen. Second, disagreements occurred during frames with a lot of movement, generated either by the child or by the caregiver repositioning the child. In one additional instance (the video in Fig. 2), other children were present in the video and the infant's face was partially occluded, resulting in labels of the other children's looks as opposed to the infant's looks.

Far generalization to entirely held-out online data set

So far, our results show that when iCatcher+ is trained on a specific data set, it can generalize to held-out videos from that same data set. How would the model perform on an entirely new data set with no further retraining? To answer this question, we conducted a test of far generalization as a proxy for what developmental researchers can expect if they use iCatcher+ on their own online data set. We took the network that was trained on the Lookit videos and, with no further training, ran inference on a different set of webcam videos (hereafter referred to as Zoom). The Zoom data set consisted of 63 videos from infants ages 7 to 10 months that roughly matched the demographics of the participants in the Lookit data set (74% White, 26% other), collected via synchronous video conferencing and recorded at a higher resolution than the Lookit videos (1280 × 720 pixels for Zoom vs. 640 × 480 pixels for Lookit). The goal of the Zoom study was to evoke different degrees of habituation and dishabituation of looking times as a target for computational modeling (Cao et al., 2022). In this study, participants saw six pairs of familiarization and test events. Each familiarization consisted of a presentation of a visual stimulus at the center of the screen for varying durations, and the test event consisted of a presentation of either the previously shown familiarization stimulus or a completely new stimulus. A naive human rater annotated looking times for each trial. For details about the annotation scheme for the Zoom data set, see Table S4 in the Supplemental Material (coding procedures available at https://osf.io/yqr6b).

Despite differences in the research topic (intuitive physics in Lookit, habituation/dishabituation in Zoom), primary dependent measure (preferential looking in Lookit, looking duration in Zoom), and primary body posture (infants held over their caregivers' shoulders in Lookit, infants sat on their caregivers' laps or in a high chair in Zoom; 57% high chair, 40% lap, 3% other), the model trained only on the Lookit videos performed well on the Zoom data set (Table 2). Although iCatcher+ had never seen a video from the Zoom data set before evaluation, it achieved average human–model frame-by-frame agreement of 85.87% (95% CI = [84.31%, 87.31%]), roughly equal to the same metric in the Lookit videos. Most importantly, iCatcher+ produced trial-level looking times, the dependent variable used in this experiment, that were comparable with human-generated looking times (Fig. 7a; ICC = 0.97, 95% CI = [0.97, 0.98]) and comparable with performance on the same measure in the Lookit test set (Table 2; ICC = 0.95, 95% CI = [0.93, 0.97]). Although high overall agreement could, in principle, hide failure to perform well on a small fraction of videos, instead we found that iCatcher+ showed good correspondence with human annotations across all videos in the Zoom data set (Fig. 7b). This suggests that the pretrained Lookit model (available at https://github.com/yoterel/icatcher_plus) can be used with no further retraining for automated, reliable annotation of new videos collected via Lookit or over live video conferencing, especially for studies that use looking duration as the primary dependent measure.

Discussion

Developmental psychology aspires to build and test theories of the mind by studying infants and young children. In the past decade, the field has developed techniques for collecting data faster than ever before,

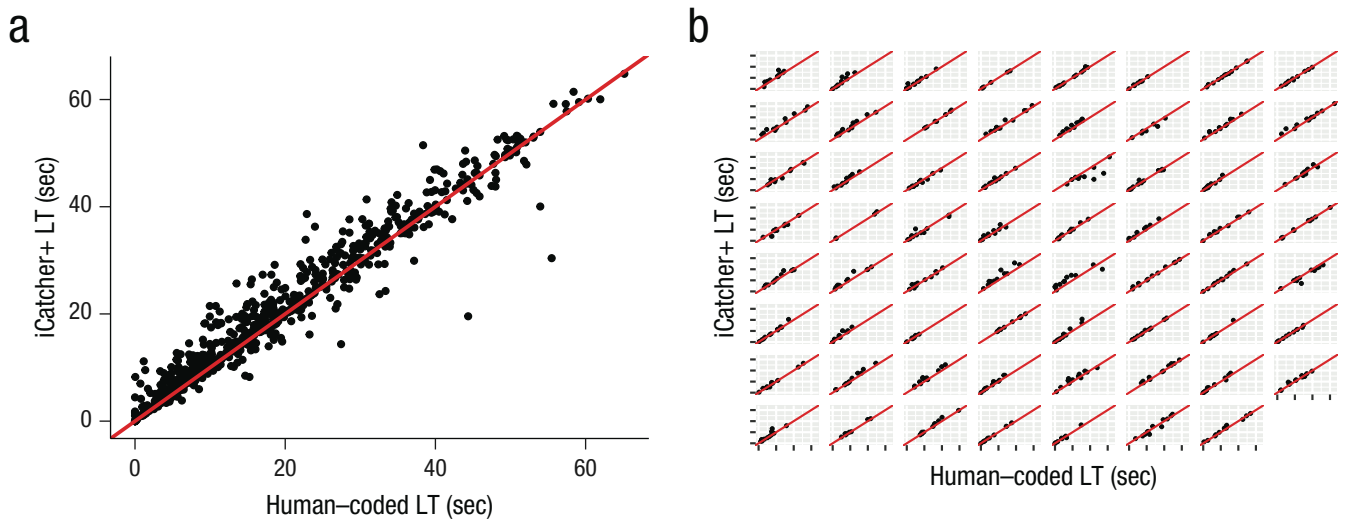

Fig. 7. Correspondence between trial-level looking time (LT) generated by a human rater versus iCatcher+ in the Zoom data set (a) across all videos and (b) broken down per video. Note that this comparison was conducted on mutually valid frames (i.e., excluding frames in which iCatcher+ could not detect a face and frames that a human rater deemed unusable for reasons such as caregiver interference, webcam lag, and experimenter error).

potentially from a vastly larger and broader set of participants. Our toolkit has expanded to include online testing (Chuey et al., 2021; Scott & Schulz, 2017), large-scale replication (Frank et al., 2017), and meta- and mega-analysis of existing data (Koile & Cristia, 2021; Tsuji et al., 2017). Developmental psychology has also partnered with the field of computational cognitive science to formalize theories about knowledge and learning and design tests of those theories (Lake et al., 2017; Shu et al., 2021; Smith et al., 2019; Tenenbaum et al., 2011). But for the field to make the most of these tools, researchers need a faster way of annotating looking behavior—the primary measure in developmental behavioral studies—from video, including videos collected online. In this article, we built on iCatcher (Erel et al., 2022), a system for gaze classification previously trained and tested on one data set collected in the lab and tested its performance on three other data sets, chosen to represent diverse research settings (university labs, in field sites, in homes) and participants (varying in age, race, and ethnicity).

Overall, we found that iCatcher+ achieved excellent performance on classifying looking behavior as LEFT versus RIGHT, with near-human frame-by-frame performance on the data sets collected in a university lab setting or a mobile lab setting in various field sites (California-BW, Senegal) and somewhat lower frame-by-frame accuracy on the data set collected online (Lookit). Even so, we consider the Lookit results to be a success because despite many sources of between-videos variability, trial-level human–model agreement, aggregated across frames, approached human performance, with room to grow. We found that performance did not vary substantially by age, gender, race and ethnicity, lighting conditions, face movement and position, and face pixel density; the single best predictor of accuracy was model confidence. The most common failure case was that iCatcher+ could not detect a face in the frame (and returned a label of INVALID), and inspection of these frames revealed that these are time points during which infants were turned away from the screen or covering their faces. Across all data sets, by swapping out these INVALID frames with the label AWAY, we found that iCatcher+ could classify ON versus OFF looks, or LEFT versus RIGHT looks, with near-human accuracy, but not both at the same time. Most compellingly, the model that was trained on the Lookit data set returned reliable annotations for a fourth, entirely held-out data set of videos from online research that it had never seen before. From these findings, we conclude that iCatcher+ meets the criteria of accuracy and robustness stipulated in the introduction, and thus we believe that this tool can be adopted by developmental psychologists to supplement or replace human annotation sometime in the near future.

Limitations and qualifications

The current research has limitations and qualifications. First, although we have shown good performance on held-out videos sampled from the same distribution as the training set (i.e., performance on Lookit videos,

given a network trained only on Lookit videos) and generalization to a separate data set (i.e., performance on Zoom data set, given a network trained on Lookit videos), this generalization surely has a limit. The high performance in the Zoom data set, although promising, could have been driven by the higher resolution of the videos, the presence of an experimenter to monitor for setup and data quality, and/or other features. We emphasize again that in the absence of random assignment across these features, it is not possible to pinpoint the cause(s) of differences or similarities in performance across the data sets presented in this article. Within each data set, we found that performance for both humans and the model was relatively robust to participant demographics, video luminance, participant pose and movement, and pixel density of the face (affected by camera resolution, viewing distance, and head size; see Figs. S6 and S7 in the Supplemental Material). However, it is plausible that there are video features that affect annotation accuracy we were unable to examine in the current article, including screen size (constant for California-BW and Senegal and not collected for Lookit or Zoom) and camera resolution (not available for California-BW or Senegal and not collected for Lookit or Zoom). We emphasize that because iCatcher+ ultimately processes images at the pixel level, developmental researchers whose setups or participants differ substantially from the data sets in this article should expect a drop in performance.

Second, our qualitative error analysis has shown that the face classifier works best when the participant's face is not covered up, in shadow, or moving quickly. Infants and toddlers cannot be instructed to avoid these behaviors. If developmental labs seek to add iCatcher+ to their research protocols, they will have to consider how to maximize the quality of their video data and whether or when human raters should be included in the process.

Third, by design, iCatcher+ was trained to classify looking behavior into three categories (LEFT, RIGHT, and AWAY) over an entire trial rather than as a continuous vector projected to a point on the screen or looking to different areas of interest over time. Thus, it is best suited for experiments that use duration looking and preferential looking as primary dependent measures. However, this framework could plausibly be extended to include more precise classification over continuous space (e.g., see Werchan et al., 2022) and time (e.g., looking behavior time-locked to the onset of a stimulus or in anticipation of an outcome). We leave this direction to future research and welcome contributions to the open-source codebase (https://github.com/yoterel/icatcher_plus).

Ongoing and future work

Fine-tuning to calibration frames before annotation. When humans annotate videos of infant looking behavior, they are instructed to inspect a series of calibration frames in which the participant's gaze was attracted to different locations on the screen or on and off the screen. With compelling enough calibration stimuli, the direction of an infant's gaze can be reliably expected to fall on specific locations on the screen at specific time intervals. Human raters use these calibration frames to make judgments about edge cases (e.g., In Frame X, is the infant looking below or at the bottom of the screen?). Currently, iCatcher+ does not take advantage of this strategy—the model is not provided with any information about the specific video to be annotated before classification. However, future versions of the model could leverage the technique of network personalization in which a pretrained model is fine-tuned using a few frames to specialize the model for that particular participant and research setting. This process involves adjusting the weights of the pretrained model in response to frames from the to-be-classified video. Park et al. (2019) introduced FAZE (few–shot adaptive gaze estimation), a framework using model-agonistic metalearning (Finn et al., 2017) that takes this approach and reports promising results on videos of adults. The authors reported boosts in accuracy of gaze classification faces that the network was not trained on when the network was first fine-tuned to a small (fewer than nine) number of randomly selected frames of the new face. In the future, we can imagine adding metalearning to the iCatcher+ pipeline in which the model is either given random frames or high-quality calibration frames from a new video before classification.

Building out the tool. Although the current iCatcher+ codebase is open source and freely available at https://github.com/yoterel/icatcher_plus, it still has a way to go before it is a user-friendly tool. Command-line interfaces can present an obstacle for researchers with less technical knowledge and experience, and as the tool continues to evolve, distributing releases over github likewise requires skills that some researchers may not possess. Thus, we are currently working on designing a web app that will be accessible to researchers regardless of technical training.

Integration with Lookit. So far, Erel et al. (2022) and the current article have shown that iCatcher+ can be adopted as a tool for annotating videos of gaze behavior once these videos have been collected. However, as mentioned in the introduction, iCatcher can also run in

"online mode," wherein the model classifies gaze behavior from an incoming video stream in real time. In Erel et al., the authors reported performance using this "live" mode, which classified a video frame every ≈42 ms (or 24 frames per second, close to the actual frame rate of the videos). Our tests have shown iCatcher+ can process each frame faster, roughly every ≈22 ms (45 frames per second) using a midtier GPU (NVIDIA GeForce RTX 2060) or every ≈59 ms (17 frames per second) without one. This can allow integration into Lookit or other frameworks for fully automated data collection and enable experimental designs, including infant-contingent stimulus presentation, that are currently not possible to run on Lookit. Although fully automated online data collection of gaze behavior in infants and children remains a distant goal, the current work represents a critical move forward.

Considerations for developmental researchers interested in using iCatcher+

Overall, we believe that iCatcher+ can be adopted by developmental psychologists in the very near future to supplement or replace human annotation for many research programs. It performs with near-human accuracy, and its failure modes and confidence scores are interpretable: iCatcher+ reports when it cannot detect a face in general or an infant face specifically and provides lower confidence scores for frames it is likely to misclassify. We envision a pipeline of automated annotation that takes a video and an event file including time stamps and types of each trial (e.g., expected, unexpected, target object on the left) and then returns (a) frame-by-frame (or time bin-by-time bin) annotations of gaze behavior (either LEFT/RIGHT/AWAY, or ON/OFF, and NOFACE, NOBABYFACE) and (b) confidence scores (0–1) for each annotation. Developmental-psychology labs can then take this information and design appropriate protocols for interpreting and analyzing these data.

iCatcher+ will not eliminate all barriers to analyzing infant and toddler gaze data, but we expect that most labs will be able to adapt their practices to smoothly transition to machine labeling. Many of these changes can be initiated immediately: For instance, labs can modify their data-collection protocols and instructions to minimize the chances that infants' faces will be partially or fully occluded (a key failure mode of iCatcher+). Second, labs should consider stimulus-presentation methods that automatically produce event time stamps. Tools with these capacities include jsPsych (de Leeuw, 2015), pyHab (Kominsky, 2019), Psychtoolbox (Borgo et al., 2012), and Lookit (Scott & Schulz, 2017). Lookit users should consider feeding each trial of data to the model individually rather than a video of concatenated trials because of drops in performance when the model averages discontinuous frames from different trials. Third, labs should consider how they would interpret iCatcher+ output, potentially in combination with human ratings. For example, a lab protocol could stipulate that a human rater go back to annotate trials for which the mean iCatcher+ confidence score falls below 0.85 or a face was not found for more than half the frames. The protocol could also define thresholds for excluding trials or participants from further analysis using these metrics.

Finally, when possible, labs should explicitly evaluate the trade-offs of time and expense for hand-coding data versus running more participants to compensate for the potential added noise of machine labeling. Erel et al. (2022) showed that it is possible to replicate a key result from an LWL study, originally generated from human-rated video data, using iCatcher. In the current work, the human-level ICC scores for both dependent measures (preferential looking and duration looking) for held-out videos from the California-BW, Senegal, and Lookit data sets, and for the entirely held-out Zoom data set, suggest that researchers should expect iCatcher+ to be as accurate as a trained human rater across trials but worse than a human rater at the level of frames. However, the impact of noise (from human or machine raters) will vary across phenomena and paradigms. Because there is currently no consistent reporting of effect sizes or standards for reporting the reliability of LT data in published developmental research, it is difficult to predict how many additional participants should be run for a study using iCatcher+ annotation for a given effect or method. However, it may be more efficient to collect and automatically label data from more infants than it would be to hand-code a smaller sample. With iCatcher+ and these new research protocols in hand, this framework can enable rapid, adequately powered research into the origins of the human mind for all developmental scientists.

Transparency

Action Editor: Pamela Davis-Kean
Editor: David A. Sbarra
Author Contribution(s)

Yotam Erel: Conceptualization; Formal analysis; Methodology; Project administration; Software; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing.

Katherine Adams Shannon: Conceptualization; Data curation; Funding acquisition; Project administration; Supervision; Writing – original draft; Writing – review & editing.

Junyi Chu: Data curation; Formal analysis; Writing – original draft; Writing – review & editing.

Kim Scott: Conceptualization; Data curation; Funding acquisition; Investigation; Project administration; Supervision; Writing – review & editing.

Melissa Kline Struhl: Data curation; Funding acquisition; Methodology; Project administration; Supervision; Validation; Writing – original draft; Writing – review & editing.
Peng Cao: Formal analysis; Methodology; Software; Validation; Visualization; Writing – review & editing.
Xincheng Tan: Formal analysis; Methodology; Software; Validation; Visualization; Writing – review & editing.
Peter Hart: Formal analysis; Methodology; Software; Validation; Visualization; Writing – review & editing.
Gal Raz: Data curation; Formal analysis; Validation; Visualization; Writing – original draft; Writing – review & editing.
Sabrina Piccolo: Data curation; Formal analysis; Writing – review & editing.
Catherine Mei: Data curation; Formal analysis; Writing – review & editing.
Christine Potter: Conceptualization; Investigation; Methodology; Supervision; Writing – review & editing.
Sagi Jaffe-Dax: Formal analysis; Methodology; Software; Supervision; Validation; Visualization; Writing – review & editing.
Casey Lew-Williams: Conceptualization; Funding acquisition; Investigation; Methodology; Supervision; Writing – original draft; Writing – review & editing.
Joshua Tenenbaum: Conceptualization; Funding acquisition; Supervision; Writing – review & editing.
Katherine Fairchild: Conceptualization; Data curation; Funding acquisition; Methodology; Project administration; Resources; Software; Supervision; Writing – review & editing.
Amit Bermano: Formal analysis; Methodology; Software; Supervision; Validation; Visualization; Writing – review & editing.
Shari Liu: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

We gratefully acknowledge the following funding sources: National Institute of Health (F32HD103363 to S. Liu, F32HD105705 to K. A. Shannon, R01HD095912 to C. Lew-Williams), Defense Advanced Research Projects Agency Machine Common Sense Program (CW3013552), Siegel Family Endowment (S4881), the Simons Foundation, the Princeton Data-Driven Social Sciences Initiative, Len Blavatnik and the Blavatnik Family Foundation, the Yandex Foundation, and the Massachusetts Institute for Technology Quest for Intelligence.

Open Practices

This article has received the badges for Open Data and Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iDs

Katherine Adams Shannon https://orcid.org/0000-0003-4069-9576
Kim Scott https://orcid.org/0000-0002-8356-3129
Sagi Jaffe-Dax https://orcid.org/0000-0002-8759-6980
Shari Liu https://orcid.org/0000-0002-7037-5401

Acknowledgments

We thank Anne Fernald for sharing video data and Virginia Marchman and the teaching team for the Spring 2021 course, "6.819/6.869: Advances in Computer Vision" at the Massachusetts Institute for Technology, for project support and guidance; Rebecca Saxe, Emily Chen, and the Cambridge Writing Group for helpful discussion and feedback; and Loey Bull, Faith Choe, Beyza Ciftci, Alice Wang, Jessica Zhu, Elisa Dimagiba, Lucy Fu, Ivy Huang, Katherine Johnson, Liora Jones, Gabriel Kane, Naomi Kirimi, Ashley Lederman, Christina Lee, Cynthia Lei, Crystal Liu, Claire Ma, Asmita Mittal, Christopher Montejo, Salina Musyaju, Alison Plump, Zoe Price, Sofia Riskin, Lillian Switkes, Khaled Shehada, Julius Tao, Kevin Wen, Anna Wilson, Jessica Zheng, Grace Zhang, and research assistants at the Language Learning Lab and Early Childhood Cognition Lab for research assistance, including assistance with video annotation. Videos from the Lookit data set with permission granted for scientific use are available at https://osf.io/5u9df/. Access to raw video files from the California Black and White Video and Senegal data sets is not available because of restricted participant privacy agreements. To protect participant privacy, participant identifiers for the video and demographic data are not linked to each other. However, this information is available upon reasonable request to Katherine Adams Shannon (kat.adams@stanford.edu).

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/25152459221147250

Notes

1. The Lookit sample contained many fewer unique children (83 vs. 214 in California-BW vs. 143 in Senegal), thus we sampled a greater proportion to ensure sufficient variability in the training set.
2. In the Lookit data set, 11 files required changes to the trial or looking-behavior labels (e.g., missing "instructions," extra "left," frame numbers contained a letter); 12 files were missing annotation of final "end," so we needed to verify total number of frames; and one annotation file was discarded and reannotated by a new person because it was missing annotations for a majority of the trials. In the California-BW data set, one file required a change in label because of a typo, and a small number of files required the manual addition of a new time stamp to indicate the start time of the experiment. In the Senegal data set, all video files were trimmed by 2 s to 11 s to sync the first frame of the time stamp with the first frame of the video. Although human annotators for all data sets were well trained, this process revealed to

us an even greater need for automated gaze annotation, which minimizes opportunity for human error.

References

Aslin, R. N. (2007). What's in a look? *Developmental Science*, *10*(1), 48–53.

Aslin, R. N., & Smith, L. B. (1988). Perceptual development. *Annual Review of Psychology*, *39*, 435–473.

Baillargeon, R., Scott, R. M., & Bian, L. (2016). Psychological reasoning in infancy. *Annual Review of Psychology*, *67*(1), 159–186.

Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L.-P. (2018). OpenFace 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)* (pp. 59–66). IEEE.

Borgo, M., Soranzo, A., & Grassi, M. (2012). Psychtoolbox: Sound, keyboard and mouse. In M. Borgo, A. Soranzo, & M. Grassi (Eds.), *MATLAB for psychologists* (pp. 249–273). Springer.

Bradski, G. (2000). The openCV library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, *25*(11), 120–123.

Byers-Heinlein, K., Bergmann, C., & Savalei, V. (2022). Six solutions for more reliable infant research. *Infant and Child Development*, *31*(5), Article e2296. https://doi.org/10.1002/icd.2296

Cao, A., Raz, G., Saxe, R., & Frank, M. C. (2022). *Habituation reflects optimal exploration over noisy perceptual samples*. PsyArXiv. https://doi.org/10.31234/osf.io/jb7qy

Chouinard, B., Scott, K., & Cusack, R. (2019). Using automatic face analysis to score infant behaviour from video collected online. *Infant Behavior & Development*, *54*, 1–12.

Chuey, A., Asaba, M., Bridgers, S., Carrillo, B., Dietz, G., Garcia, T., Leonard, J. A., Liu, S., Merrick, M., Radwan, S., Stegall, J., Velez, N., Woo, B., Wu, Y., Zhou, X. J., Frank, M. C., & Gweon, H. (2021). Moderated online data-collection for developmental research: Methods and replications. *Frontiers in Psychology*, *12*, Article 4968. https://doi.org/10.3389/fpsyg.2021.734398

Cubuk, E. D., Zoph, B., Shlens, J., & Le, Q. V. (2020). Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 702–703). IEEE.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). IEEE.

Erel, Y., Potter, C. E., Jaffe-Dax, S., Lew-Williams, C., & Bermano, A. H. (2022). iCatcher: A neural network approach for automated coding of young children's eye movements. *Infancy*, *27*(4), 765–779. https://doi.org/10.1111/infa.12468

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314.

Fernald, A., & Marchman, V. A. (2012). Individual differences in lexical processing at 18 months predict vocabulary growth in typically developing and late-talking toddlers. *Child Development*, *83*(1), 203–222.

Fernald, A., Pinto, J. P., Swingley, D., Weinberg, A., & McRoberts, G. W. (1998). Rapid gains in speed of verbal processing by infants in the 2nd year. *Psychological Science*, *9*(3), 228–231.

Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language comprehension by infants and young children. *Developmental Psycholinguistics: On-Line Methods in Children's Language Processing*, *190*, 97–135.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 1126–1135). PMLR.

Fischer, T., Chang, H. J., & Demiris, Y. (2018). RT-GENE: Real-time eye gaze estimation in natural environments. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 334–352). Springer

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., Hamlin, J. K., Hannon, E. E., Kline, M., Levelt, C., Lew-Williams, C., Nazzi, T., Panneton, R., Rabagliati, H., Soderstrom, M., Sullivan, J., Waxman, S., & Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421–435.

Friedman, S. (1972). Newborn visual attention to repeated exposure of redundant vs "novel" targets. *Perception & Psychophysics*, *12*(3), 291–294.

Haith, M. M. (1980). *Rules newborns look by*. Erlbaum.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*(7169), 557–559.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). IEEE.

Horowitz, F. D., Paden, L., Bhana, K., & Self, P. (1972). An infant-control procedure for studying infant visual fixations. *Developmental Psychology*, *7*(1), 90. https://doi.org/10.1037/h0032855

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization*. arXiv. https://doi.org/10.48550/arXiv.1412.6980

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, *83*(2), B35–B42.

Koile, E., & Cristia, A. (2021). Toward cumulative cognitive science: A comparison of meta-analysis, mega-analysis, and hybrid approaches. *Open Mind: Discoveries in Cognitive Science*, *5*, 154–173.

Kominsky, J. F. (2019). PyHab: Open-source real time infant gaze coding and stimulus presentation software. *Infant Behavior & Development*, *54*, 114–119.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *The Behavioral and Brain Sciences*, *40*, Article e253. https://doi.org/10.1017/S0140525X16001837

Lew-Williams, C., & Fernald, A. (2007). Young children learning Spanish make rapid use of grammatical gender in

spoken word recognition. *Psychological Science*, *18*(3), 193–198.

Lukyanenko, C., & Fisher, C. (2016). Where are the cookies? Two- and three-year-olds use number-marked verbs to anticipate upcoming nouns. *Cognition*, *146*, 349–370.

Marchman, V. A., & Fernald, A. (2008). Speed of word recognition and vocabulary knowledge in infancy predict cognitive and language outcomes in later childhood. *Developmental Science*, *11*(3), F9–F16.

Oakes, L. M. (2012). Advances in eye tracking in infancy research. *Infancy: The Official Journal of the International Society on Infant Studies*, *17*(1), 1–8.

Oakes, L. M. (2017). Sample size, statistical power, and false conclusions in infant looking-time research. *Infancy*, *22*(4), 436–469. https://doi.org/10.1111/infa.12186

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). WebGazer: Scalable webcam eye tracking using user interactions. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 3839–3845). IEEE.

Park, S., Mello, S. D., Molchanov, P., Iqbal, U., Hilliges, O., & Kautz, J. (2019). Few-shot adaptive gaze estimation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9368–9377.

Parker, R. A., Scott, C., Inácio, V., & Stevens, N. T. (2020). Using multiple agreement methods for continuous repeated measures data: A tutorial for practitioners. *BMC Medical Research Methodology*, *20*(1), Article 154. https://doi.org/10.1186/s12874-020-01022-x

Roberts, S. O., Bareket-Shavit, C., Dollins, F. A., Goldie, P. D., & Mortenson, E. (2020). Racial inequality in psychological research: Trends of the past and recommendations for the future. *Perspectives on Psychological Science*, *15*(6), 1295–1309.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Scott, K., & Schulz, L. (2017). Lookit (Part 1): A new online platform for developmental research. *Open Mind*, *1*(1), 4–14.

Shu, T., Bhandwaldar, A., Gan, C., Smith, K. A., Liu, S., Gutfreund, D., Spelke, E., Tenenbaum, J. B., & Ullman, T. D. (2021). *AGENT: A benchmark for core psychological reasoning*. arXiv. https://doi.org/10.48550/arXiv.2102.123

Simion, F., Regolin, L., & Bulf, H. (2008). A predisposition for biological motion in the newborn baby. *Proceedings of the National Academy of Sciences, USA*, *105*(2), 809–813.

Slater, A., Morison, V., & Rose, D. (1984). Habituation in the newborn. *Infant Behavior & Development*, *7*(2), 183–200.

Smith, K., Mei, L., Yao, S., Wu, J., Spelke, E., Tenenbaum, J., & Ullman, T. (2019). Modeling expectation violation in intuitive physics with coarse probabilistic object representations. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper/2019/hash/e88f243bf341ded9b4ced444795c3f17-Abstract.html

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review*, *99*(4), 605–632.

Teller, D. Y. (1979). The forced-choice preferential looking procedure: A psychophysical technique for use with human infants. *Infant Behavior & Development*, *2*, 135–153.

Telles, E. (2014). *Pigmentocracies: Ethnicity, race, and color in Latin America*. UNC Press Books.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, *58*(1), 267–288.

Tsuji, S., Bergmann, C., Buckler, H., Cusack, R., & Zaadnoordijk, L. (2021). *Toward a large-scale collaboration for infant online testing: Introducing ManyBabies-AtHome*. Max Planck Institute for Psycholinguistics. https://www.mpi.nl/publications/item3281985/toward-large-scale-collaboration-infant-online-testing-introducing

Tsuji, S., Bergmann, C., Lewis, M., Braginsky, M., Piccinini, P., Frank, M. C., & Cristia, A. (2017). MetaLab: A repository for meta-analyses on language development, and more. In *Interspeech* (pp. 2038–2039). https://www.isca-speech.org/archive_v0/Interspeech_2017/pdfs/2053.PDF

Valenza, E., Simion, F., Cassia, V. M., & Umiltà, C. (1996). Face preference at birth. *Journal of Experimental Psychology. Human Perception and Performance*, *22*(4), 892–903.

Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognitive Psychology*, *29*(3), 257–302.

Werchan, D. M., Thomason, M. E., & Brito, N. H. (2022). OWLET: An automated, open-source method for infant gaze tracking using Smartphone and webcam recordings. *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-022-01962-w

Xu, F., Carey, S., & Welch, J. (1999). Infants' ability to use object kind information for object individuation. *Cognition*, *70*(2), 137–166.

Zielinski, P. (2007). *Opengazer: Open-source gaze tracker for ordinary webcams*. Samsung and The Gatsby Charitable Foundation. http://www.Inference.Phy.Cam.Ac.Uk/opengazer