This dataset was created to study the use of aesthetic manipulation in cookie banners. Aesthetic manipulation is a type of dark design pattern that can generally be described as a design choice in a user interface intended to distract users' attention from one thing and/or direct their attention towards something else [1]. In the case of cookie banners, it is a design choice that is made to draw a user's attention to the "Accept" button, which allows companies to collect and process the consenting user's personal data. This datasheet is created based on the guidelines of [2].

For questions or concerns about the published dataset, contact Riley Grossman (rag24@njit.edu). This dataset can be repurposed for any future academic research so long as the original paper is cited. Several potential uses of this dataset are listed in the README file of this repository.

There are five datasets released as part of this project. We refer to them throughout the rest of the datasheet as follows:

1. Image dataset – refers to the stored images of website landing pages that were collected for 2,492 websites from a NY IP address and 2,490 websites from an EU IP address.
2. D1- refers to the "image_labels.csv" dataset that has manually labeled categories assigned to each of the images in the image dataset.
3. D2- refers to the "labeled_banners.csv" file that contains the manually labeled bounding box coordinates for each of the 1,623 cookie banners in the image dataset that give users a choice to opt-out of data collection/processing, and the accept button, reject button, manage button, and close button on each of these. The bounding box coordinates are used to aggregate all pixel-level salience scores to button-level scores (based on maximum and average values).
4. D3- refers to the "button_salience_characteristics.csv" which contains the summed salience score for all 4,610 buttons belonging to the 1,621 images containing a cookie banner with 2+ buttons in the dataset, along with button characteristics.
5. D4- refers to the "robust_salience_scores.csv" file, which has aggregated button-level salience scores for 32 perturbed versions of each of the 894 images of cookie banners that are compliant, and display accept, reject, and manage buttons.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

The images dataset is comprised of images. These images are screenshots of the webpage (a.k.a. landing page) a user will see when first loading each of the websites in the dataset for the first time (i.e., no previous visits). D1, D2, D3, and D4 contain data that is processed after manually labeling (and altering in the case of D4), the images in the images dataset.

How many instances are there in total (of each type, if appropriate)?

There are 4,982 images in the images dataset. D1 has 2,956 website domains listed. Each website has 2 labels (5,912 labels) because the websites are visited from EU and NY IP addresses. There are more labels than images because 930 of the webpage visits failed (e.g., IP address is blocked or webpage is broken). D2 has the manually obtained bounding box information and button-level salience data for 1,623 of the images in the images dataset. It specifically has bounding boxes for the images of websites implementing a cookie banner that complies with GDPR by allowing the user a choice to opt-out of data processing/collection. D3 has characteristics for 4,610 buttons on the 1,621 images from D2 that have 2+ buttons. D4 has 28,608 total observations. It is comprised of button-level salience data for 32 perturbed versions of each of the 894 images of websites implementing a compliant cookie banner with accept, reject, and manage buttons.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The images dataset is just a sample of all possible websites. We collected the top 1,000 websites according to the Tranco list [3], as well as 1,000 randomly selected websites from outside the top 1,000. Furthermore, we selected the top 500 websites with a ccTLD in the EU, and then 500 randomly ranked ones from outside the top 500. Due to 44 overlapping websites in the top 1,000 total websites and top 500 EU websites, there are 2,956 websites selected in total. The D1-D4 datasets are samples of the image datasets, as explained in previous answer.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

The images dataset is raw images of the landing pages of websites.

D1 has the domain of each website ("Domain"), the Tranco rank ("Rank"), how it was selected to the sample ("Type") (e.g., Top 500 EU websites), a binary indictor for whether there was a DNS issue loading the website which makes it unreachable ("DNS_Issue"), and the manually labeled image category based on whether the website is reachable, whether it has a banner, and which of the 17 designs the banner falls under, if existing. The manually labeled image is broken into two variables based on whether the website was visited from an NY ("Label_final_NYIP") or EU IP address ("Label_final_EUIP"), as the implementation is often different.

D2 has the domain ("shortened_url"), whether ccTLD is in the EU ("EU"), whether it was visited from EU IP address ("EUIP"), and the manual label from D1 ("Final_label"). It also has the manually labeled bounding box for the cookie banner, and each of its buttons stored as dictionary ("label"). It also has a manually created list of which buttons are hidden in the text or formatted as a link ("Note"). The area of the cookie banner and each button (if available) are recorded ("Banner", "X" for close button, "Accept" , "Reject", and "Manage"). For each of these five elements, the average pixel-level salience score of all pixels making it up is recorded, as well as the maximum pixel-level salience score of all of these pixels ("____salience_avg" and "____salience_max" respectively). Finally, the average and maximum salience level score of all pixels in the image is stored ("image_salience_avg" and "image_salience"max").

For each button instance recorded in D3, the image label is recorded as [insert website domain]_[Boolean value for if it is in the EU] ("URL") and the cookie banner label from D1 is recorded under "label". For each button, the type is stored (e.g., accept or reject) under "button", and the average and maximum salience scores from D2 are normalized between 0-1 and summed together ("salience"). The size of the cookie banner ("banner_size") and distance between center of the banner and center of the page are recorded ("banner_distance"). Similarly, absolute and relative button size ("button_size" and "button_size_p"), and the distance from the center of the button to the center of the page ("button_distance") and center of the cookie banner are recorded ("button_banner_distance"). Each button's average greyscale brightness ("brightness"), the banner's greyscale brightness ("banner_brightness"), and absolute difference between the two is recorded ("contrast"). Finally, Boolean values are stored to indicate

whether each button is hidden in the corner (“corner”) or text (“hidden”), and/or formatted as a link (“link”) or choice menu (“choices”) instead of a button.

D4 contains the same information as D2, but for perturbed images. However, it does not store the image label, “Note”, “image_salience_max”, “image_salience_avg”, or any of the button areas.

Is there a label or target associated with each instance? If so, please provide a description.

The image dataset could be used for classification with D1’s “Label_final_NYIP” and “Label_final_EUIP”, or object identification with D2’s “label” bounding boxes used as the target.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Some instances in D2 may not contain all button options. In this case, the values for average and max salience, as well as area, are left as 0.

Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)? If so, please describe how these relationships are made explicit.

Relationships are made explicit between datasets because the URL (or domain name) will identify all instances relating to a particular image. The EUIP value indicates if the image is collected from EU or NY IP address.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please describe these splits, explaining the rationale behind them.

No splits are used in the paper. If doing so, images from the same website but different IP addresses should be placed in the same dataset split. Otherwise, basically identical images could end up in both training and test splits.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Some images could be redundant if they look the exact same when visiting from EU and NY (see above).

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

Self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why. If the dataset does not relate to people, you m

No.

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The images are directly observable and collected (via screenshots) with a Python-based crawler. The cookie banner labels, bounding boxes, and several button characteristics (e.g., link/hidden) are manually derived from the original images. The salience scores are derived from a salient object detection model called DeepRare (which takes images as input). The remaining data points are derived from the manual labels and DeepRare outputs.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

The Selenium package in Python is commonly used for building simple web crawlers. We did not perform our own validation the Selenium web driver as it is used as part of OpenWPM, a widely used, open-source tool for web privacy measurement [4]. For manual labeling, we validated our approach by having three people manually label a subset of 200 cookie banner images. The three raters had high interrater reliability

(Krippendorff's alpha of 0.94). We also used the unsupervised, generalizable salient object detection model DeepRare [5]. The creators of DeepRare show that the model is designed for application to new image datasets and that its unsupervised performance is more generalizable than other unsupervised salient object detection models. We did not have access to ground-truth salience maps for cookie banner images. Thus, we chose DeepRare because it generalizes well to unseen and diverse types of images.

If the dataset is a sample from a large set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The image dataset is a sample from all possible websites we could visit. It is partially deterministic (i.e., top 500 EU websites and top 1,000 global websites) and partially randomly samples (i.e., 500 EU websites outside the top 500 and 1,000 global websites outside the top 1,000). D1 is a label for each website chosen I the images dataset. D2 is a deterministically sampled subset of the image dataset. It has instances for every image that contains a cookie banner giving the user a real choice to opt-out of data collection (e.g., a reject or manage settings button). D3 has all of the same instances as D2. D4 is a deterministically sampled subset of D2 containing all of the cookie banner images deemed to have compliant designs and accept, reject, and manage buttons.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The data was collected between December 16th and 23rd, 2024. The instances captured were the real-time implementations of websites.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No.

Preprocessing, Cleaning, and Labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

The image dataset is raw, but D1-D4 all involve processing and labeling.

D1 was manually labeled. We reviewed cookie banners and discovered that we could group their designs into 17 mutually exclusive and exhaustive categories. We determined key features that determined the design, which are based on which buttons are presented on the first page of the banner. We used these developed criteria to label each cookie banner into one of the 17 categories.

D2 requires further manual labeling to denote the bounding boxes around each button on the cookie banners, and identify if buttons are formatted as a link or hidden in the text. The bounding boxes can then be used to determine the size of each button, as well as to aggregate the pixel-level salience score outputs from DeepRare into button-level salience scores. We process the images by inputting them to DeepRare to create pixel-level salience maps. These maps are then used to create the button-level salience scores.

D3 requires some further processing to collect button characteristics (e.g., button brightness). These steps can be explored in detail in the "Explaining_Button_Salience.ipynb" file in the repository.

D4 involves more preprocessing to perturb each input image slightly. The perturbed images are again input to DeepRare and we use the output salience maps to calculate more robust measurements of button salience.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

We store the images dataset, which is the basic raw data which all further datasets are based on. We further save the manually labeled bounding box raw data in D2 so that it can be used for tasks other than calculating button characteristics and salience scores. We did not provide all of the pixel-level salience maps used in D2 and D4 (over 30,000 images total), but we provide a description of how they can be generated. If necessary, future researchers may contact us, and we can find a way to provide access.

Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

We used Label Studio to provide bounding boxes around each cookie banner and button. It is available: https://labelstud.io/guide/.

[1] Gray, Colin M., et al. "The dark (patterns) side of UX design." *Proceedings of the 2018 CHI conference on human factors in computing systems*. 2018.

[2] Gebru, Timnit, et al. "Datasheets for datasets." *Communications of the ACM* 64.12 (2021): 86-92.

[3] Pochat, Victor Le, et al. "Tranco: A research-oriented top sites ranking hardened against manipulation." *arXiv preprint arXiv:1806.01156* (2018).

[4] Englehardt, Steven, and Arvind Narayanan. "Online tracking: A 1-million-site measurement and analysis." Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. 2016.

[5] Kong, Phutphalla, et al. "DeepRare: generic unsupervised visual attention models." Electronics 11.11 (2022): 1696.