DATASHEET: AVALON

This document is based on *Datasheets for Datasets* by Gebru *et al.* [1]. Please see the most recent version here.

MOTIVATION

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

We created this RL environment in order to provide a more complex environment in which others can train more robust RL systems that are better able to generalize.

Our set of evaluation levels, and human play data recorded on them were explicitly created for the purpose of measuring baseline human performance in order to better understand and contextualize the performance of RL systems.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

This dataset was created by the team at Generally Intelligent, a research company.

What support was needed to make this dataset? (e.g.who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)

No specific additional funding was required for this dataset.

Any other comments?

No.

COMPOSITION

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description. The dataset has 3 components:

1. The RL environment. This is software which generates worlds and allows agents and human players to interact with those worlds. This is the primary contribution and focus of the work.

2. The fixed set of evaluation worlds. These worlds are Godot "scene" files that were generated by the world generation code (mentioned in the previous point). They include references to art assets, materials, and other resources that must be available in order for the scene to be loaded and interacted with (and which are included in the RL environment).

3. The set of human play data. This consists of a set of human "play throughs" of each of the evaluation worlds. At a 1/10th of a second resolution, it defines the exact movements of the VR controllers and headset, as well as other game values (ex: whether the player gained life from eating fruit or lost life from being hit by a predator).

How many instances are there in total (of each type, if appropriate)? 1. There is 1 simulator that has 20 different "tasks", each of which can generate an infinite number of worlds.

2. There are 1,000 evaluation worlds (50 for each of the 20 tasks)

3. There are 217 hours of human playthroughs on the 1,000 worlds (roughly 5 playthroughs per level).

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). 1. The RL environment itself is able to generate an unbounded number of worlds for each task. These are sampled in a biased way during our suggested training procedure, since that bias leads to significantly better baseline performance.

2. The evaluation environments were sampled randomly, with some small caveats. See Appendix G for a more in-depth discussion of the exact selection procedure.

3. The human data represents most of the human data we collected–it only excludes data on a small number of "practice" levels where we told participants that their performance would not matter. The selection of the human participants was done by asking for volunteers from friends, and so is certainly a biased. We do not believe this to be an unbiased or representative sample of the population. Even within those who could have been selected, we specifically filtered for practical concerns (people who had VR headsets and would have sufficient time to collect data over a short window).

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

1. The simulator is an OpenAI Gym environment and a set of functions that generate *.tscn files.

2. The evaluation worlds are the set of *.tscn files that were generated and used to score the baseline networks.

3. The raw data that was recorded from the players' controllers and headsets. See this function for details on the for-

mat: `get_observations_from_human_recording`

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

No.

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.

No. That said, VR controllers can occasionally lose tracking temporarily, but the game does not pause, so that simply counts as if the controllers were not moved. Even if the game crashed, data was uploaded (though it would count as an incomplete playthrough).

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.

Yes. All data for a single user is associated with a single user id.

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

Yes. 100% of the evaluation levels should be used for testing only. They should not be used for validation or for training. Data for validation and training should come from the generator using a different random seed.

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

There are certainly errors in the simulator. We maintained a bug tracker during the human data recording so that we could investigate issues as they were reported. Some issues were fixed immediately, while others required more investigation. We intend to fix all known issues before the wider public release. Where the errors or bugs interefered with the ability to play a world or the world was found to be impossible, we disabled that world, and as described in the main paper and appendix, this affected only a few percent of worlds.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset of human playthroughs is certainly more useful with the simulator, though it does not strictly rely on it.

a) Yes, these resource will continue to exist and remain constant. We have committed the code to Github, and it can be used to exactly reproduce our evaluation and represents the version that the data was collected with.

b) Yes, we've created a DOI for the dataset.

c) There are no license fees or other restrictions on our dataset. The entire artifact is open source.

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.

Not intentionally. If the dataset is viewed in VR, it can be a little bit scary to be attacked by a low-poly bear or jaguar in VR, or to look down from a high cliff for those who have a fear of heights.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Most of the work is about the simulator, which does not relate to people.

The only part that relates to people is the human playthroughs.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No. All data was anonymized.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

No.

COLLECTION

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 data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Human play data was recorded directly and uploaded to our servers.

Evaluation worlds were generated via a script which is included in the released code.

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. Finally, list when the dataset was first published.

Human play data was collected over a period of approximately 4 days.

The dataset was first published as of this submission, on June 16th, 2022.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

All human play data was collected via Meta Quest 2 VR devices and uploaded to our servers.

A small amount of manual curation was used while selecting the exact evaluation levels, as described in the paper.

The distributions of levels were validated by looking at hundreds of worlds for each task to ensure that they seemed possible and were of approximately the right level of difficulty. None of the examined worlds were used in the evaluation set.

What was the resource cost of collecting the data? (e.g. what were the required computational resources, and the associated financial costs, and energy consumption - estimate the carbon footprint. See Strubell *et al.*[2] for approaches in this area.)

The energy requirements were extremely minimal (the calories burned by the players over the course of the 217 hours of gameplay and the battery charges on each VR device, plus the energy to run the docker container hosting the server).

By far the largest direct cost was the monetary cost from compensating human participants, which worked out to roughly $12,000.

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

The evaluation worlds were sampled as described in Appendix G.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Volunteers were selected from a pool of people who expressed interest in an online form. They were compensated by being given the VR device that they used to collect the data, and by being paid $30 / hr for any time that exceeded the cost of purchasing the VR device in the first place. This was the compensation scheme that was proposed in the original signup form.

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.

Yes. We conducted an internal review process to determine the risks and best practices associated with conducting a human study of this type. We evaluated the potential risks associated with VR gameplay and made a number of recommendations for instructions and safety guidelines that went above and beyond the normal precautions for this activity. See our instructions to participants included in the supplementary materials.

Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.

Yes.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

Directly.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Yes. See supplementary material for exact wording.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Yes. Please see the Adult Consent Form in the supplementary materials.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate)

No. It is not possible to revoke their consent after this data has been published because copies of that information may have been made by third parties, and that is allowed under the license with which we have released the dataset. This was made clear in the consent form.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

Yes. The data is anonymized, and thus is at very low risk of having any significant impact on the study participants.

Any other comments?

No.

PREPROCESSING / CLEANING / 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 remainder of the questions in this section.

No.

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.

Yes. The data link contains the raw human control data, and postprocessing is not relevant for the simulator or evaluation levels.

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

NA.

Any other comments?

No.

USES

Has the dataset been used for any tasks already? If so, please provide a description.

The dataset of human playthroughs has been used to calculate average human performance. See the main paper for more details.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

No, though we may create one after release.

What (other) tasks could the dataset be used for?

Any number of RL research questions could be asked in our simulated environment.

For the human playthrough data, one could imagine using the data to understand human game play, look for training effects, or do behavior cloning, among other uses. However, we suspect there are other, more interesting uses for the data that we have not thought of, which is one reason why we wanted to release it.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

It should not be taken as in any way being representative of any larger population.

Many users had played some practice levels before, and that data is not completely recorded, which might make it more difficult to make any definitive claims about training effects.

Are there tasks for which the dataset should not be used? If so, please provide a description.

No.

Any other comments?

No.

DISTRIBUTION

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes. We selected a CC-BY-SA license expressly for the purpose of making the data easy to distribute and share.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

At the very least, the dataaset is currently hosted in a public S3 bucket. It will likely also be hosted on our website, as well as on github.

All of the code for the simulator is on github.

Yes, we have created a DOI for the dataset of human data and evaluation worlds. See the official note for a link to the dataset and DOI.

When will the dataset be distributed?

The dataset will be put online starting on June 16th, 2022 so that it can be easily reviewed, though we do not intend to promote it or widely distribute the links until Dec 1st, 2022.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

No. The license for the human playthrough dataset is a CC-BY-SA license so that others can easily re-use the data.

The simulator itself will be released under a GPL license because we want to encourage any users to contribute their changes back to the community.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

No.

MAINTENANCE

Who is supporting/hosting/maintaining the dataset?

Generally Intelligent.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The manager of the dataset can be emailed directly at kanjun@generallyintelligent.ai

Is there an erratum? If so, please provide a link or other access point.

See the github repository for a log of changes and errors as they are fixed.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

Yes. We intend this simulator to be a living artifact, continually being updated to introduce bug fixes, more tasks and configuration options, and new capabilities. We will maintain a change log with major and minor releases available on github.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

No, there are no limits on the retention of the data.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

We plan to regularly update the dataset with new fixes and features. Once released, older major versions will likely not be updated because it will make it difficult to identify and compare between subtly different versions of the same dataset. This is precisely why we plan to fully release the dataset and simulator later this year–while they are complete and functional as they are today, the additional months of polishing will allow us to create a very stable and reliable base on which other researchers can build. Future release will also likely exist in a "preview" state for a significant amount of time before being more widely released.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Yes. Others are welcome to build on the dataset and release their changes in whatever way they see fit. We will attempt to resolve issues posted on Github and merge any pull requests that would benefit other users. All pull requests will be reviewed and tested before being merged, and updated releases will be posted to github and added to the change log as described above.

Any other comments?

We plan to be responsible stewards of this dataset and simulator. We will be using it extensively internally, and plan to continually release the improvements that we make.

REFERENCES

[1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. Datasheets for Datasets. *arXiv:1803.09010 [cs]*, January 2020.

[2] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and Policy Considerations for Deep Learning in NLP. *arXiv:1906.02243 [cs]*, June 2019.