Further Details on Examining Adversarial Evaluation: Role of Difficulty

Behzad Mehrbakhsh [a,b,c;*], **Fernando Martínez-Plumed**[a,b] **and José Hernández-Orallo**[a,b,c]

[a]UPV - Universitat Politècnica de València
[b]VRAIN - Valencian Research Institute for Artificial Intelligence
[c]ValgrAI - Valencian Graduate School and Research Network of Artificial Intelligence

A Technical Appendix

This supplementary material serves as technical appendix with sections given detailed information about 1) how difficult the different datasets are; 2) the distribution of class levels within each difficulty range across all datasets; 3) an illustrative selection of images in each difficulty bin and 4) performance comparison of M_{orig} and M_{SAdv} on a per-instance basis for each difficulty level.

A.1 Class difficulty

Figure 2 shows the difficulty distribution per dataset and class, sorted by difficulty. Of course, the difficulty level of the classes for the different datasets can vary depending on the population of systems used to solve the classification tasks. Still, in our case, we are using a large population of systems (1000 for MNIST, 449 for FashionMNIST and 156 for CIFAR10), using the data from [19], which in turn comes from the OpenML platform[1].

Starting with MNIST, this is a dataset of handwritten digits from 0 to 9. The images in the MNIST dataset are normalised and centred, which makes the dataset less challenging (for image classification tasks) compared to other datasets such as CIFAR-10. In general, all the classes in the MNIST dataset are considered relatively easy to classify, as the images are well-segmented and have a clear contrast. However, some digits such as 5, 8 seem to be slightly more difficult to classify (probably because they are more similar to other digits like 6 or 9 respectively).

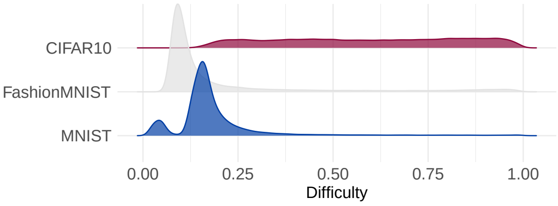

Figure 1: Difficulty distribution per dataset. Sorted by average difficulty.

Newer datasets such as FashionMNIST (a dataset of images of clothing and accessories) are considered to be more challenging than MNIST in terms of complexity and diversity. However, we see that FashionMNIST does not make a great difference in terms of the aggregated distribution of difficulty compared to MNIST (see Figure 1). In general, some classes in the FashionMNIST dataset are more difficult to classify than others. For example, the trousers, bags, sneakers or sandals classes are relatively easy to classify, while the shirt, pullover, coat or t-shirt are considered more difficult (see the tail to the right of their difficulty distributions). This is probably due to the similarities in the images of these classes and the variations in the texture, shape, and color.

Finally, we see higher average difficulties and differences between classes for CIFAR10. The images in this dataset are relatively complex, with objects that are often occluded, partially visible or in non-frontal poses. These factors make the dataset more challenging than the above datasets. What we see in Figure 2 is that, in general, the vehicle-related images for CIFAR10 are considered relatively easy, while the animal-related images are considered more difficult. This is due to the similarities in the images of these classes, which can make them hard to distinguish. Some images of deer and horse are similar and can cause confusion to a model.

A.2 Class distribution

Figure 3 shows the class distribution for each difficulty range for the original and all adversarial datasets.

For the original dataset D_{Orig}, we see that some classes are easy, and have a high proportion of very easy instances: class 1 (digit 1) for MNIST, classes 1 (trouser), 8 (bag) and 7 (sneaker) for Fashion-MNIST, and class 8 (ship) for CIFAR10. Other difficult classes are more dominated by difficult instances: class 5 (digit 5) for MNIST, class 6 (shirt) for FashionMNIST, and class 3 (cat) for CIFAR10.

For the Simple Adversarial dataset (D_{SAdv}), the number of difficult instances has increased, but in an uneven way for some classes over others (note that the y-axis changes across plots).

For the Balanced Adversarial dataset (D_{BAdv}), is constructed in the same way as D_{SAdv}, we see how they have increased the number of difficult instances for all classes, but now the number of instances for each class is the same (which does not mean of course that the distribution of difficulty is the same for all classes).

For Double Adversarial dataset (D_{DAdv}), as we undersample the easiest instances, there are no instances in the lower bins for all

[*] Corresponding Author. Email: bmehrba@upv.es.
[1] https://www.openml.org/

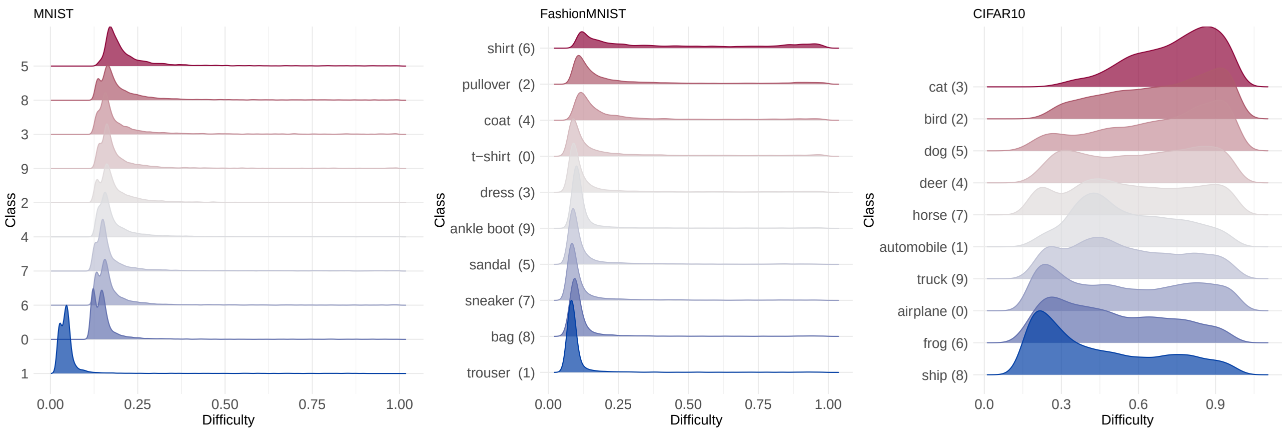

Figure 2: Difficulty distribution per dataset and class. Classes sorted by average difficulty.

classes.

A.3 Sample images for each difficulty bin

Figure 4 shows sample images from each difficulty bin for MNIST, FashionMNIST and CIFAR10. Images belonging to the first three bins are relatively easy instances, but the images of the two hardest bins (d and e) are sometimes really challenging even from a human perspective. A human will find it difficult or will misclassify the examples provided in this figure for the hardest bin (Figure 4 (e)).

A.4 M_{orig} VS. M_{SAdv} confusion matrix

Figure 5 compares the performance of M_{Orig} and M_{SAdv} on a per-instance basis for each difficulty level on the dataset constructed from MNIST, FashionMNIST and CIFAR10. We can observe that M_{Orig} performs equally good or in most cases better than M_{SAdv} in the first 4 bins when tested on MNIST and FashionMNIST and only performs worse in the last bin. We see almost the same trend for CIFAR10 with the difference of M_{SAdv} performing slightly better in the fourth bin, too.

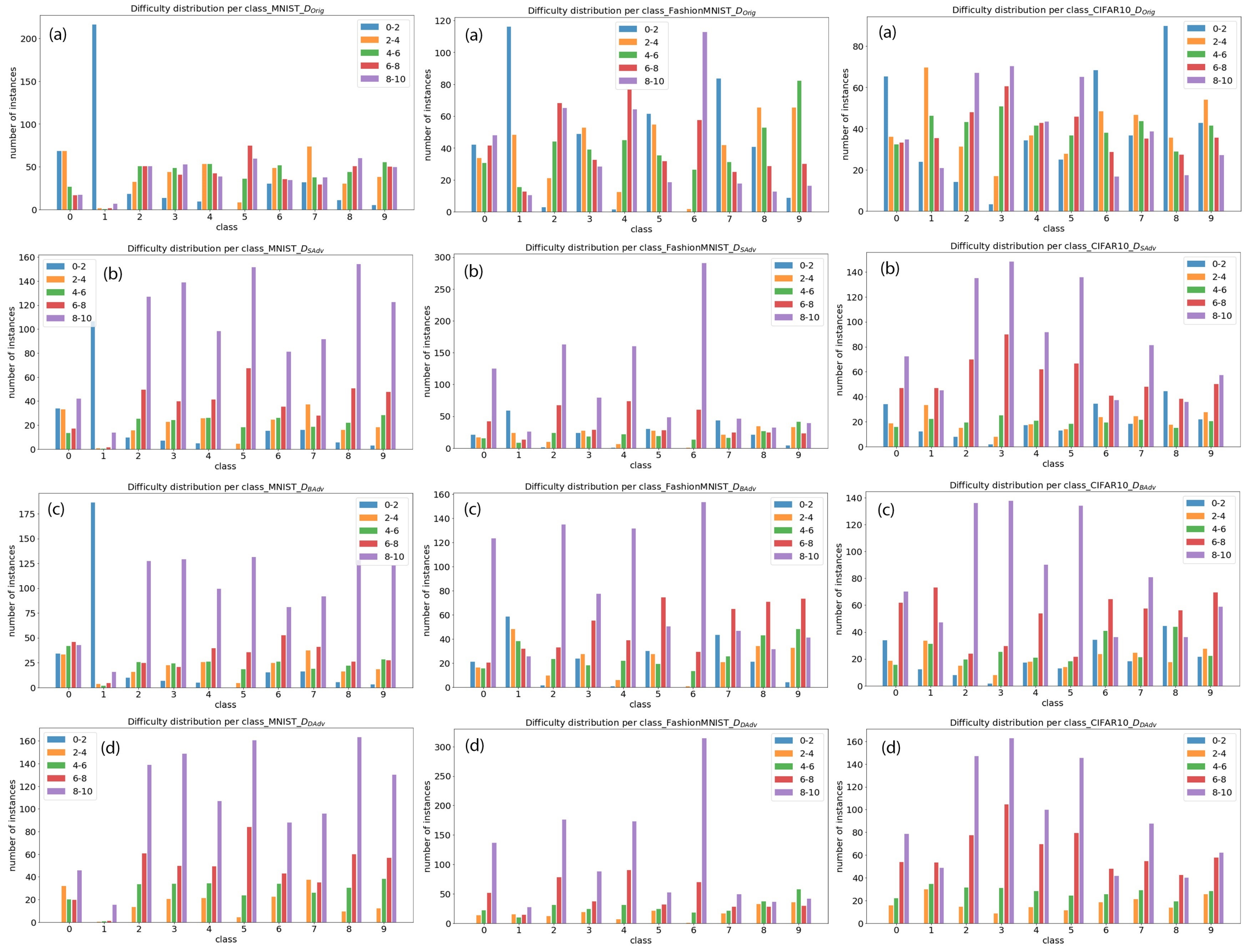

Figure 3: Proportion of each class for each difficulty range for all modified datasets constructed from MNIST (left), FashionMNIST (centre), and CIFAR10 (right), for D_{Orig} (a), D_{SAdv} (b), D_{BAdv} (c), D_{DAdv} (d). Class names for FashionMNIST and CIFAR10 in Figure 2.

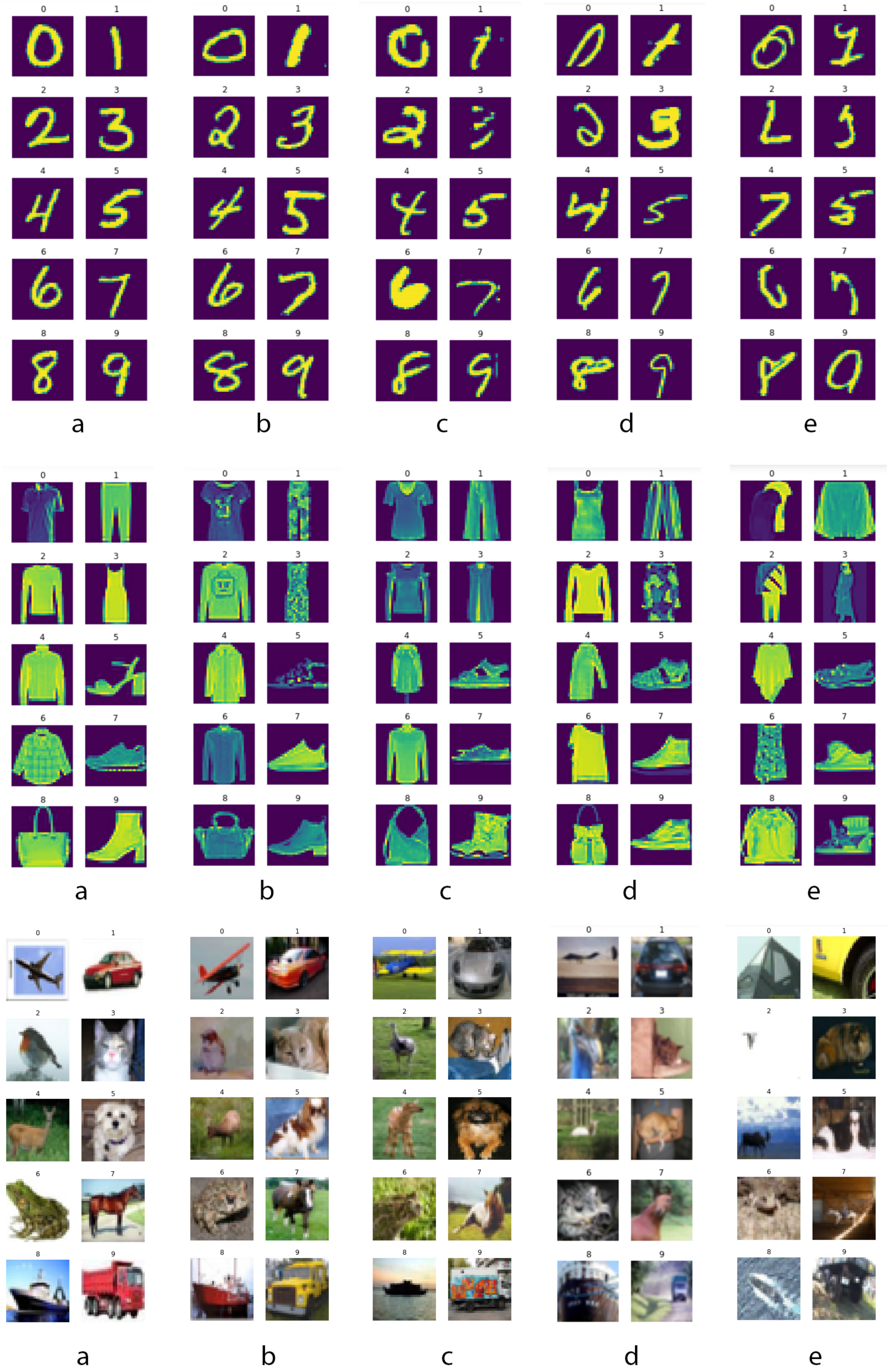

Figure 4: The figure presents sample images from the MNIST (top), FashionMNIST (centre), and CIFAR10 (bottom) datasets organised by difficulty bin. The bins are designated (a) through (e); (a) represents the easiest bin, (b) the difficulty level between 2 and 4, (c) the difficulty level between 4 and 6, (d) the difficulty level between 6 and 8, and (e) the hardest bin which covers difficulty level between 8 and 10.

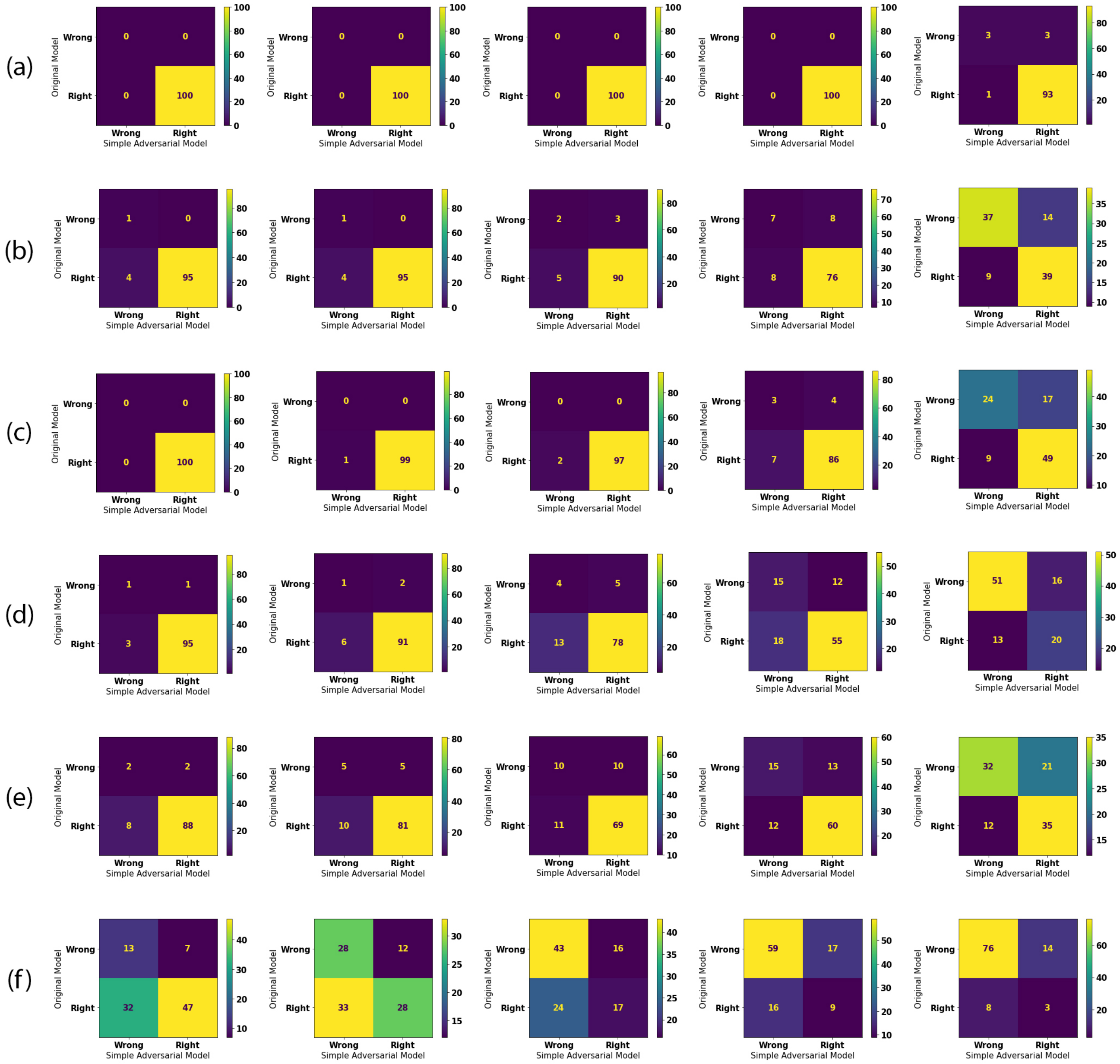

Figure 5: This graph compares the performance of M_{orig} and M_{SAdv} on a per-instance basis for each difficulty level on MNIST ((a) using CNN and (b) using NN), FashionMNIST ((c) using CNN and (d) using NN), and CIFAR10 ((e) using CNN and (f) using NN).

Acknowledgments

We thank the anonymous reviewers for their comments. This work was funded by valgrAI, the Norwegian Research Council grant 329745 Machine Teaching for Explainable AI, the Future of Life Institute, FLI, under grant RFP2-152, the EU (FEDER) and Spanish grant RTI2018-094403-B-C32 funded by MCIN/AEI/10.13039/501100011033 and by CIPROM/2022/6 funded by Generalitat Valenciana, EU's Horizon 2020 research and innovation programme under grant agreement No. 952215 (TAILOR), US DARPA HR00112120007 (RECoG-AI) and Spanish grant PID2021-122830OB-C42 (SFERA) funded by MCIN/AEI/10.13039/501100011033 and "ERDF A way of making Europe" In compliance with the recommendations of the Science paper about reporting of evaluation results in AI [3], we include all the results at the instance level.

References

[1] Martin Arjovsky, *Out of distribution generalization in machine learning*, Ph.D. dissertation, New York University, 2020.

[2] Eli Bronstein, Sirish Srinivasan, Supratik Paul, Aman Sinha, Matthew O'Kelly, Payam Nikdel, and Shimon Whiteson, 'Embedding synthetic off-policy experience for autonomous driving via zero-shot curricula', *arXiv preprint arXiv:2212.01375*, (2022).

[3] Ryan Burnell, Wout Schellaert, John Burden, Tomer D Ullman, Fernando Martinez-Plumed, Joshua B Tenenbaum, Danaja Rutar, Lucy G Cheke, Jascha Sohl-Dickstein, Melanie Mitchell, Douwe Kiela, Murray Shanahan, Ellen M. Voorhees, Anthony G. Cohn, Joel Z Leibo, and Jose Hernandez-Orallo, 'Rethink reporting of evaluation results in AI', *Science*, **380**(6641), 136–138, (2023).

[4] Brandon Carter, Siddhartha Jain, Jonas W Mueller, and David Gifford, 'Overinterpretation reveals image classification model pathologies', *Advances in Neural Information Processing Systems*, **34**, 15395–15407, (2021).

[5] Rafael Jaime De Ayala, *Theory and practice of item response theory*, Guilford Publications, 2009.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, 'Imagenet: A large-scale hierarchical image database', in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, (2009).

[7] Desmond Elliott, 'Adversarial evaluation of multimodal machine translation', in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2974–2978, (2018).

[8] S. E. Embretson and S. P. Reise, *Item response theory for psychologists*, L. Erlbaum, 2000.

[9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, 'Explaining and harnessing adversarial examples', *arXiv:1412.6572*, (2014).

[10] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song, 'Natural adversarial examples', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, (2021).

[11] José Hernández-Orallo, Marco Baroni, Jordi Bieger, Nader Chmait, David L Dowe, Katja Hofmann, Fernando Martínez-Plumed, Claes Strannegård, and Kristinn R Thórisson, 'A new AI evaluation cosmos: Ready to play the game?', *AI Magazine*, **38**(3), (2017).

[12] Robin Jia and Percy Liang, 'Adversarial examples for evaluating reading comprehension systems', *arXiv preprint arXiv:1707.07328*, (2017).

[13] Anjuli Kannan and Oriol Vinyals, 'Adversarial evaluation of dialogue models', *arXiv preprint arXiv:1701.08198*, (2017).

[14] Divyansh Kaushik, Douwe Kiela, Zachary C Lipton, and Wen-tau Yih, 'On the efficacy of adversarial data collection for question answering: Results from a large-scale randomized study', *arXiv preprint arXiv:2106.00872*, (2021).

[15] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al., 'Dynabench: Rethinking benchmarking in nlp', *arXiv preprint arXiv:2104.14337*, (2021).

[16] Alex Krizhevsky. Learning multiple layers of features from tiny images. https://www.cs.toronto.edu/~kriz/cifar.html, 2009.

[17] Kyungmin Lee, Hunmin Yang, and Se-Yoon Oh, 'Adversarial training on joint energy based model for robust classification and out-of-distribution detection', in *2020 20th International Conference on Control, Automation and Systems (ICCAS)*, pp. 17–21. IEEE, (2020).

[18] Fernando Martínez-Plumed, Pablo Barredo, Sean O Heigeartaigh, and José Hernández-Orallo, 'Research community dynamics behind popular ai benchmarks', *Nature Machine Intelligence*, **3**(7), 581–589, (2021).

[19] Fernando Martínez-Plumed, David Castellano, Carlos Monserrat-Aranda, and José Hernández-Orallo, 'When AI difficulty is easy: The explanatory power of predicting IRT difficulty', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7719–7727, (2022).

[20] Fernando Martínez-Plumed and José Hernández-Orallo. AI results for the Atari 2600 games: difficulty and discrimination using IRT. Evaluating General-Purpose Artificial Intelligence, August 20, 2017, 2nd Intl. Workshop held in conjunction with IJCAI, Melbourne, 2017.

[21] Fernando Martínez-Plumed and José Hernández-Orallo, 'Dual indicators to analyse AI benchmarks: Difficulty, discrimination, ability and generality', *IEEE Transactions on Games*, **12**(2), 121–131, (2020).

[22] Fernando Martínez-Plumed, Ricardo B. C. Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo, 'Making sense of item response theory in machine learning', in *ECAI 2016 - 22nd European Conference on Artificial Intelligence, Best Paper Award*, pp. 1140–1148, (2016).

[23] Fernando Martínez-Plumed, Ricardo BC Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo, 'Item response theory in AI: Analysing machine learning classifiers at the instance level', *Artificial Intelligence*, **271**, 18–42, (2019).

[24] Jason Phang, Angelica Chen, William Huang, and Samuel R Bowman, 'Adversarially constructed evaluation sets are more challenging, but may not be fair', *arXiv preprint arXiv:2111.08181*, (2021).

[25] Pranav Rajpurkar, Robin Jia, Percy Liang, and ., 'Know what you don't know: Unanswerable questions for squad', *arXiv preprint arXiv:1806.03822*, (2018).

[26] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar, 'Do imagenet classifiers generalize to imagenet?', in *International Conference on Machine Learning*, pp. 5389–5400. PMLR, (2019).

[27] Laasya Samhita and Hans J Gross, 'The "clever hans phenomenon" revisited', *Communicative & integrative biology*, **6**(6), e27122, (2013).

[28] David Schlangen, 'Language tasks and language games: On methodology in current natural language processing research', *arXiv preprint arXiv:1908.10747*, (2019).

[29] Michael R Smith, Tony Martinez, and Christophe Giraud-Carrier, 'An instance level analysis of data complexity', *Machine learning*, **95**(2), 225–256, (2014).

[30] Damien Teney, Ehsan Abbasnejad, Kushal Kafle, Robik Shrestha, Christopher Kanan, and Anton Van Den Hengel, 'On the value of out-of-distribution testing: An example of goodhart's law', *Advances in Neural Information Processing Systems*, **33**, 407–417, (2020).

[31] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry, 'From imagenet to image classification: Contextualizing progress on benchmarks', in *International Conference on Machine Learning*, pp. 9625–9635. PMLR, (2020).

[32] Eric Wallace, Adina Williams, Robin Jia, and Douwe Kiela, 'Analyzing dynamic adversarial training data in the limit', *arXiv preprint arXiv:2110.08514*, (2021).

[33] Eric Wallace, Adina Williams, Robin Jia, and Douwe Kiela, 'Analyzing dynamic adversarial training data in the limit', in *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 202–217, Dublin, Ireland, (May 2022). Association for Computational Linguistics.

[34] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman, 'Superglue: A stickier benchmark for general-purpose language understanding systems', *arXiv preprint arXiv:1905.00537*, (2019).

[35] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi, 'Hellaswag: Can a machine really finish your sentence?', *arXiv preprint arXiv:1905.07830*, (2019).

[36] Daniel Zhang, Saurabh Mishra, Erik Brynjolfsson, John Etchemendy, Deep Ganguli, Barbara Grosz, Terah Lyons, James Manyika, Juan Carlos Niebles, Michael Sellitto, et al., 'The ai index 2021 annual report', *arXiv preprint arXiv:2103.06312*, (2021).