Trustworthy Adaptation of DNN Classifiers

Abanoub Ghobrial[1], Hamid Asgari[2], Kerstin Eder[1]

[1]Trustworthy Systems Lab, Department of Computer Science, University of Bristol, UK
[2]Technology and Innovation Research, Thales, Reading, UK

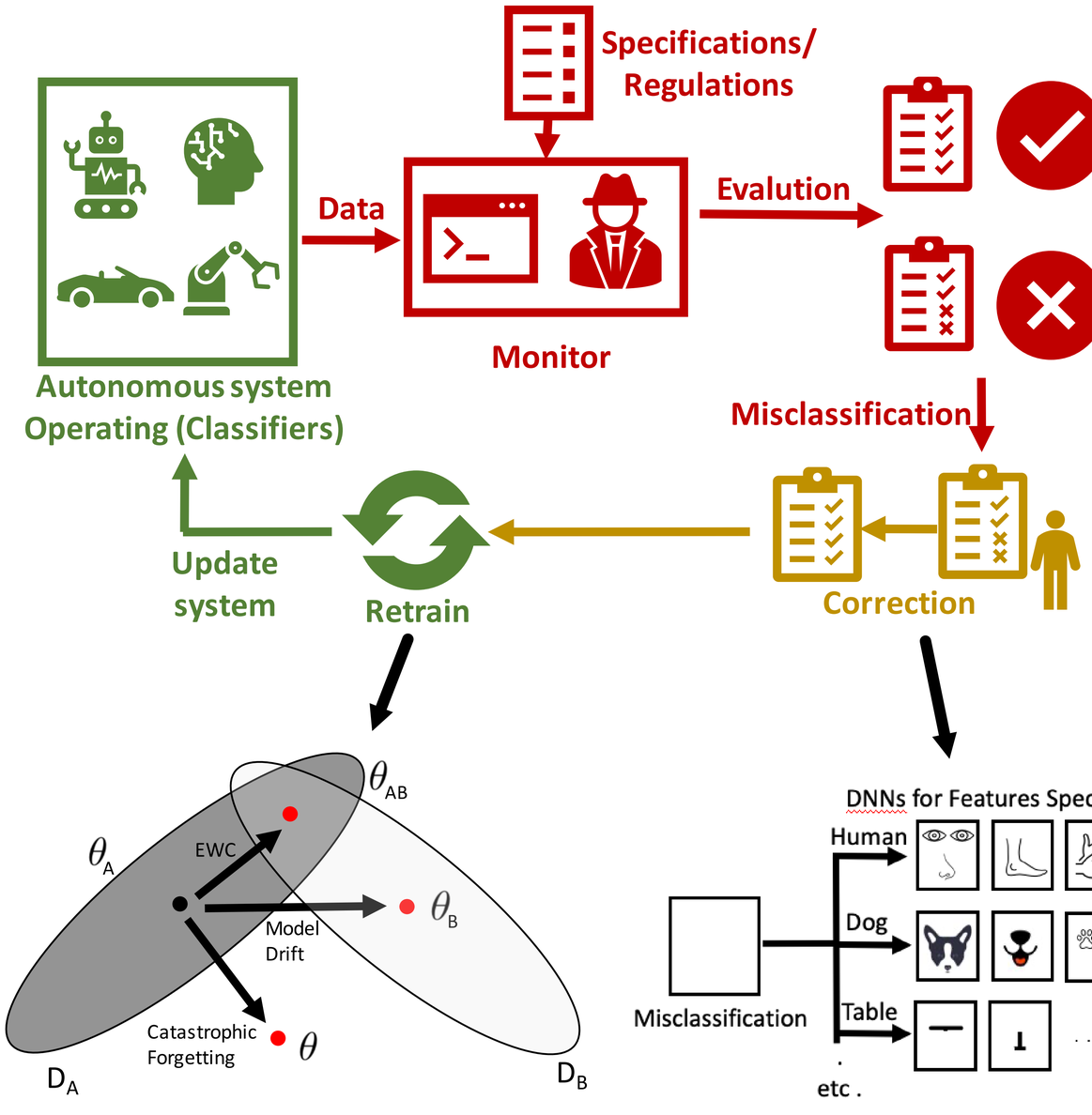

Fig 4. Retraining using EWC [3].

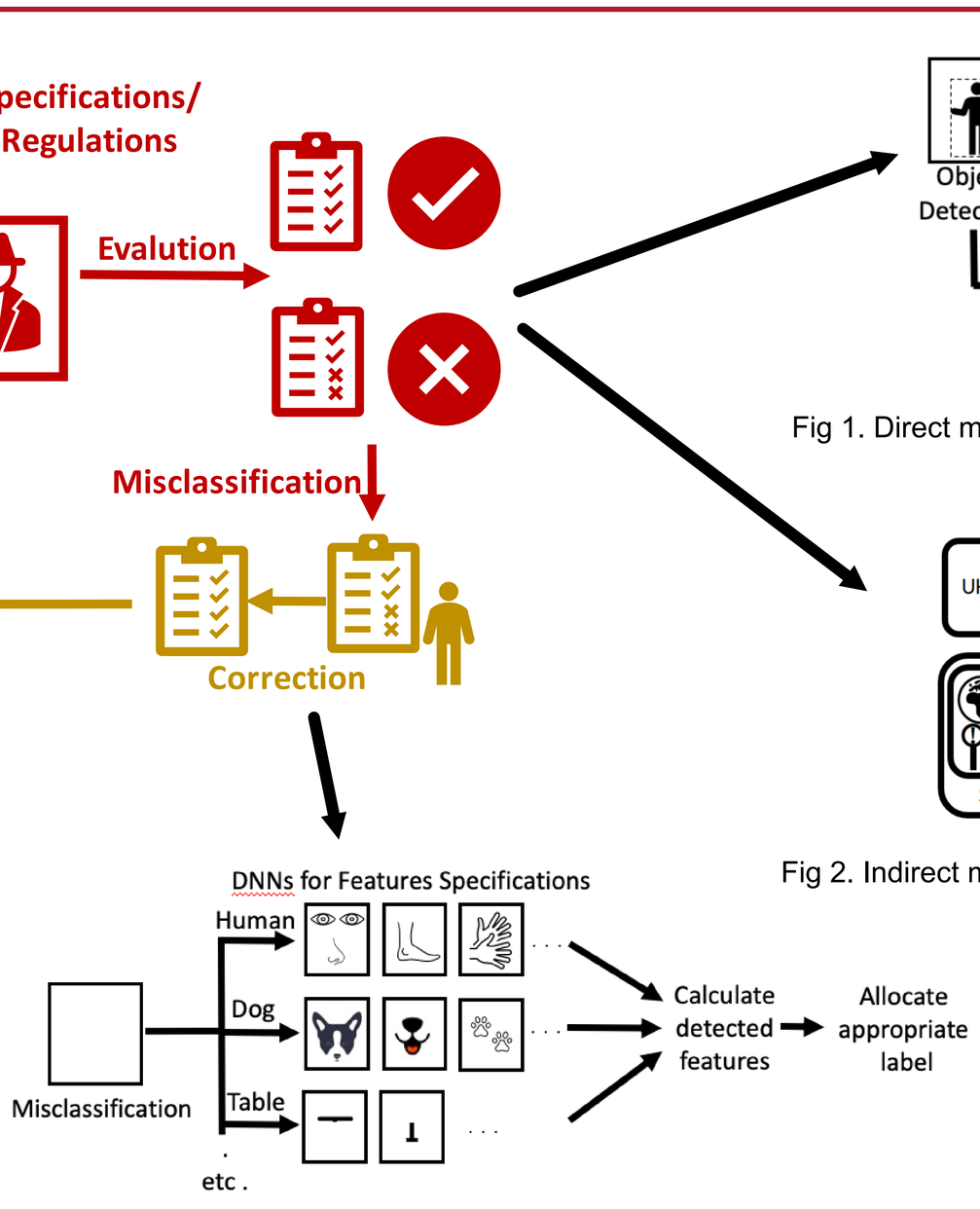

Fig 3. Concept for automated label correction.

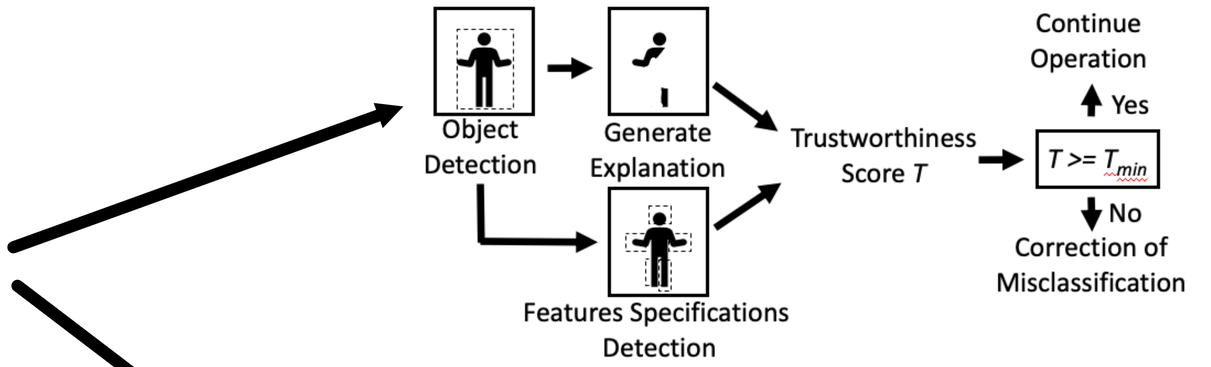

Fig 1. Direct misclassification detection using Trustworthiness Score [1].

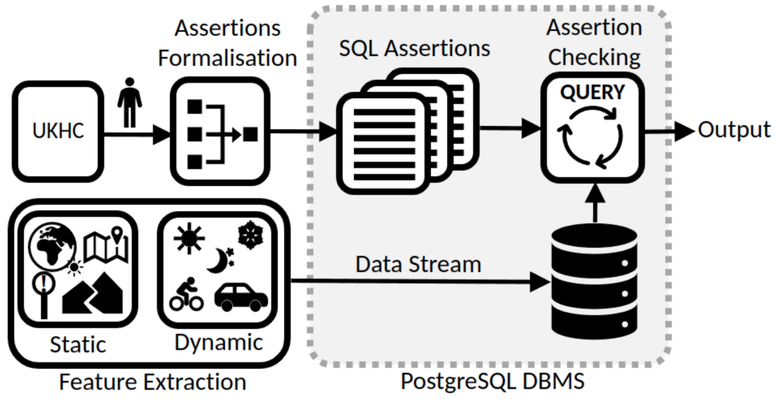

Fig 2. Indirect misclassification detection using Assertions Monitoring [2].

References:

[1] Ghobrial, A., Asgari, H. and Eder, K., 2022. Misclassifications Detection using a Measure of Trustworthiness. arXiv preprint.

[2] Harper, C., Chance, G., Ghobrial, A., Alam, S., Pipe, T. and Eder, K., 2021. Safety Validation of Autonomous Vehicles using Assertion-based Oracles. arXiv preprint arXiv:2111.04611.

[3] Ghobrial, A., Zheng, X., Hond, D., Asgari, H. and Eder, K., 2022. Operational Adaptation of DNN Classifiers using Elastic Weight Consolidation. arXiv preprint arXiv:2205.00147.