Verification Staircase: a Design Strategy for Actionable Explanations

Martin Lindvall[*]
Center for Medical Image Science and Visualization
Linköping University, Sweden
martin@ixd.ai

Jesper Molin
Sectra AB
Linköping, Sweden
jesper.molin+iui@gmail.com

ABSTRACT

What if the trust in the output of a predictive model could be acted upon in richer ways than a simple binary decision of accept or reject? Designing assistive AI tools for medical specialists entails supporting a complex but safety-critical decision process. It is common that decisions in this domain can be decomposed to a combination of many smaller decisions. In this paper, we present Verification Staircase – a design strategy that can be used for such scenarios. The verification staircase is when multiple interactive assistive tools are combined to allow for a nuanced amount of automation to aid the user. This can support a wide range of prediction quality scenarios, spanning from unproblematic minor mistakes to misleading major failures. By presenting the information in a hierarchical way, the user is able to learn how underlying predictions are connected to overall case predictions, and over time, calibrate their trust so that they can choose the appropriate level of automatic support.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; *Interface design prototyping*; *User interface design*; *User centered design*; • **Computing methodologies** → *Machine learning*; • **Social and professional topics** → Automation.

KEYWORDS

human-in-the-loop systems, human-ML collaboration, explanations, interaction design

ACM Reference Format:
Martin Lindvall and Jesper Molin. 2020. Verification Staircase: a Design Strategy for Actionable Explanations. In *Proceedings of the IUI workshop on Explainable Smart Systems and Algorithmic Transparency in Emerging Technologies (ExSS-ATEC'20).* Cagliari, Italy, 6 pages.

[*] Also with Sectra AB.

1 INTRODUCTION

Machine learning (ML) techniques has potential impacts on clinical decision making in the field of digital pathology, however, a barrier is adapting experimental results into everyday clinical use. One issue is that while results in experimental settings show impressive overall results, there is usually a relevant subset of cases where the model performs significantly worse than humans [6, 8]. Other issues such as dataset shift [7] and bias [13] also motivate a model of interaction with the predictive component positioned in the loop of human decision making.

In this paper, based on our experiences from a dual industrial-academic perspective, we outline a design strategy that we believe can be useful to resolve some of the challenges with designing human-ML collaborative systems.

The primary issues addressed by our proposed design strategy is enabling the user to answer questions such as:

- When do I trust the prediction enough to use automatic support, and when should I employ another diagnostic method?
- How can I justify my decision if a colleague asks?
- How can I feel safe in my conclusion?

In our suggested design strategy, multiple characteristics combine to enable answers to such questions, including in-the-loop correction, decomposition to allow explanations through causal inference and designing to afford use with both high performing predictions as well as border-case accuracies.

Our insights are from ongoing human-centered design explorations. The presented perspective is rooted in our experience as UX practitioners within the field of digital pathology, with a strong emphasis on practical relevance. Typically, the goal of our design effort is to make systems where the resulting value is co-created between artifacts and humans in the context of use [1]. Thus we approach explainability pragmatically, starting from users' goals and needs. Our account is less concerned about taxonomy such as distinguishing between explanations, justifications, interpretability and transparency [5] and more on our goal of creating systems that in a near future could aid clinicians to create better patient outcomes.

The layout of this paper is as follows; first we present and motivate the strategy of verification staircase. Second, we illustrate the concept by an explorative design case for assisted quantification in digital pathology. Finally, we discuss our concept in the context of explainable intelligent user interfaces and outline our proposed continuation of the research.

2 FROM CLIFFS TO STAIRCASES

Consider a predictive model trained to assess whether a patient is eligible to receive some cancer-inhibiting drug. In the context of digital pathology, where tissues are viewed at high magnification, the result might be visualized in the context of the area of interest as depicted in Figure 1.

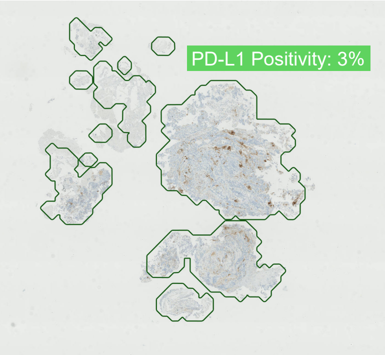


Figure 1: A diagnostic recommendation by a predictive model presented in the context of a digital pathology image

For such an interaction, the user is supposed to look at the visualization and if everything looks fine, accept the overall result. An appropriate strategy might be to trust and accept the result if the underlying accuracy is good enough for this particular case and reject it otherwise. If the user rejects the result, they will need to resort to performing the task manually. If the user interface affords no other means of judging the underlying accuracy than the manual approach, chances are that unless there exist very strong guarantees that the model performs well on all possible cases, they will always reject the result and be forced to perform their manual method.

We call this kind of human-ML interaction a verification cliff, as depicted in Figure 2

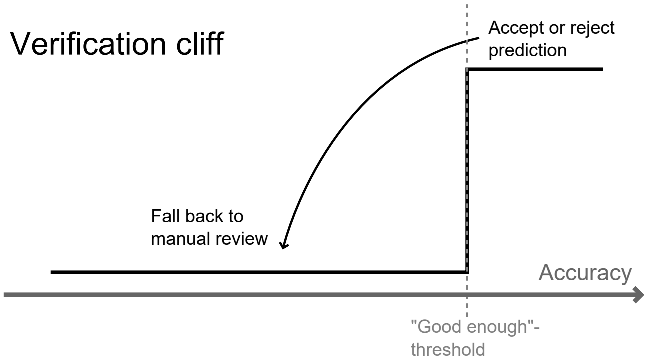


Figure 2: In some human-ML interfaces the user must decide to either accept or reject the prediction based on their belief about the underlying accuracy

What if there instead were multiple levels at which human-ML collaboration could be performed? Having modes of human operation corresponding to nuanced levels of control have long been recognized as important factors for interaction with automation [10, 11].

We argue that the performance characteristics of many ML applications make them suitable for splitting collaboration into several levels, in a similar manner to the hierarchies of ecological interface design [14]. We will next illustrate this for our pathology scenario.

Many diagnostics tasks within pathology can be divided into multiple sub tasks, e.g. an overall case-level score is derived from a formula combining the detection and classification of many individual cells. Consequently, it is possible to measure the accuracy per diagnostic case. When predictive algorithms are evaluated, it is common that an overall accuracy across cases in the form of an AUC, F1-score or Cohen's kappa is presented. However, in a scenario with case-level sub tasks, we can also characterize the distribution of per case accuracies over a large number of cases, see Figure 3

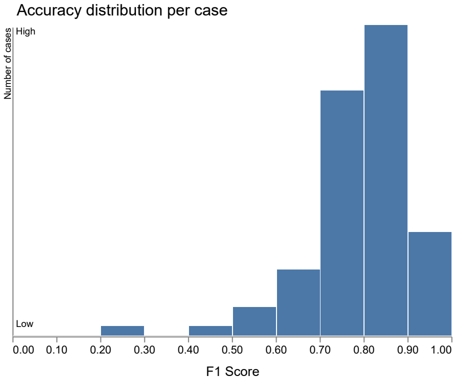


Figure 3: Common distribution of per case accuracies of a predictive model. There is a peak close to the average accuracy for the test set and then a long tail of cases.

The shape that is seen in the figure is typical and has been observed for many applications in our research. There is usually a peak in the distribution corresponding to the average accuracy and then a long tail of cases, with some cases almost always completely failing.

A better design strategy would be to think about how we can help our users when predictions fall within different intervals on the distribution. It can in many cases be possible to divide the design into multiple interactions, such as:

(1) A good result visualization that can be used to quickly verify predictions on the 0.9-1.0 span
(2) A correction tool for small modifications of predictions that updates the overall result on the 0.7-0.9 span
(3) A semi-automatic aid not even based on the original predictions on the 0.4-0.7 span etc.

This way we could attempt to create multiple user interfaces aimed at helping the user when predictions happen to fall in different positions on the accuracy distribution.

The decision of whether to trust or not trust the prediction would now be a question of *degree* - the placed trust could guide the choice to an interaction with an appropriate level of automatic support. The question then becomes: How would the user learn in which level to place their trust?

We suggest that requiring that levels are connected, correctable and composable together with visualizations that make errors apparent, could be enough. In such a design, users should be able to dynamically move between interaction levels and perform corrections. Actions at one level should immediately be reflected in the others. We argue that this combination of actionable and composable levels will enable users to calibrate their trust over time, through learning to correlate top-level observations with the suitable amount of drill-down behavior. We call this strategy a verification staircase, as depicted in Figure 4.

In the following part of this paper we will describe an ongoing case study where we have instantiated this design strategy for a tool that aids quantification in digital pathology.

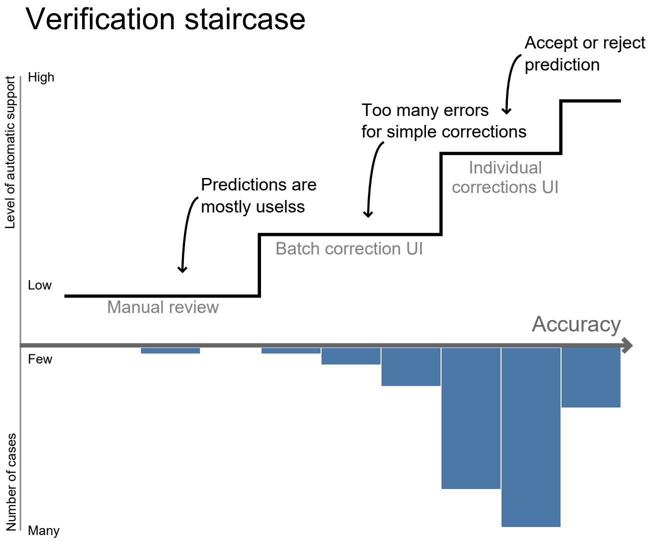


Figure 4: In a verification staircase, multiple assistive interactions are combined in a way such that the user can move between them. While some levels mean more work and less support, they give more control and a better understanding of the predictions on lower-level phenomena. Corrections at lower levels affect higher levels, and vice versa. Each level is designed to allow the human-AI ensemble to be productive within an interval of (imperfect) prediction accuracy.

3 DESIGNING WITH THE STAIRCASE: ASSISTED QUANTIFICATION

Method

We followed an iterative user-centered design (UCD) methodology combining sketching, high fidelity (hi-fi) prototyping, data collection, model debugging, user observations and interviews. Pathologists and clinical experts were consulted throughout the process. Compared to traditional UCD, we used hi-fi prototyping earlier and more frequently. This is motivated by the difficulty of eliciting how the predictive output will be experienced and behave through sketches and other low fidelity methods. Our account of the design process selectively highlights those insights we believe are important for appropriation and adaptation of the concept of verification staircase to other domains.

Diagnostic task

The assisted quantification task targeted in our case study is to determine the ratio of two types of cells. Some cancers hide from the immune system by a kind-of cloaking mechanism and can effectively be treated by disabling the cancerous cells' ability to do this. However, not all cancers hide by this mechanism. In order to determine whether a patient shall receive this expensive treatment, cells are stained such that the cell membrane of cells having the cloaking ability becomes brown. According to the diagnostic protocol, for treatment to be effective more than 50% of the cancerous cells in the tissue should have a stained membrane. If the tissue has more than 1% stained cells, the treatment might be effective. If stained cells are below 1%, the treatment will likely not work, and the patient should not be offered the treatment.

Thus, the diagnostic decision is based on estimating or counting this ratio in a possibly large tissue area. This task can be time-demanding and error-prone. Pathologists can use two basic strategies; they can look at the overall impression of the image and use their experience and tacit knowledge to "intuitively" determine the percentage right away. This is a very fast decision but can be error-prone. The second strategy involves manually counting tumor cells both with and without stained membrane, and then deriving the ratio of the two. All things being equal, this second method will result in a more accurate decision but is orders of magnitude more time-demanding. As a middle ground, pathologists sometimes choose a much smaller area as a "representative sample", and only count within that area.

A machine learning-based predictive model has the potential to always use the second strategy, classifying at the cell level and reporting the exact ratio deriving from the two counts.

Design process

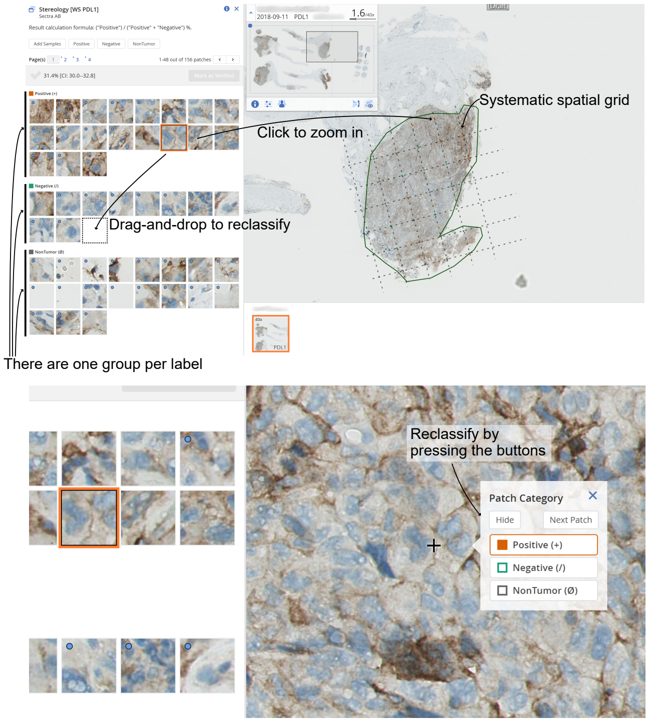
We interviewed and observed the working processes of pathologists performing the task manually. We also reviewed the available diagnostic protocols, where available. We collected and manually annotated cases and then trained a convolutional deep neural network to perform the predictions.

In one possible interaction, the user can delineate an area and receive the final result of the model as a percentage, as was depicted in Figure 1. The type of this interaction is the verification cliff – the user has two options; either they accept the result blindly or they reject it and perform their usual manual procedure. Based on the notions of a verification staircase, we sought another interaction where, if the user does not accept the top-most level of automation, they could step down to a lower level of automatic support, that is still easier and faster than manual work.

We designed our first intermediate level for the case when most cells have received the correct classification, but a few need to be corrected for a satisfactory overall result. In the devised interface, the user can explore the top-level prediction by viewing and verifying a systematic subset of decisions of the underlying model, as depicted in Figure 5.

At this level, the user is presented with a gallery of patches, sampled in a systematic spatial grid, where the patches are visually grouped according to whether they are considered to represent stained cells or not. For verification, the user can click a patch to review it in full magnification. The user can reclassify a patch by buttons in the magnified view or by drag-and-drop in the gallery. As soon as the user changes a patch, the final (top-level) ratio is updated (e.g. `31.4% [CI 30.0 - 32.8]`).

We considered showing the decision of the model for each and every pixel point (a "heatmap"), but this does not fulfill our criteria for the composability of levels. Verifying and correcting every pixel-level decision would be unfeasible for most humans. In order to not create a barrier to the higher level of the summative cell ratio, we thus limit the output of the model to grid-sampled patches. The percentage is always reported with a calculated confidence interval, reflecting the uncertainty derived from only making decisions on a subset of the tissue's cells.



Stereology [WS PDL1]
Sectra AB

Result calculation formula: ("Positive") / ("Positive" + "Negative") %.

Page(s) 1 2 3 4 1-48 out of 156 patches

31.4% [CI: 30.0--32.8]

Figure 5: The UI for the first intermediate level focused on individual classifications. The predictions are patches sampled in a grid (top right) and can be interacted with either in a gallery of patches (left) or in the context of the tissue (right). The user is able to correct false classifications by clicking and dragging in the gallery, or by clicking a patch in the magnified main view (bottom, right).

To support cases where the ratio is very close to a decision cut-off, the user can increase the certainty of their decisions by adding patches, making the sampling grid denser.

In evaluation with pathologists, we found that while this design was useful for a large subset of clinical cases where the diagnosis was far from a decision cut-off, there existed cases where the needed precision created a grid so dense that the amount of verification overwhelmed the user, and again they had to resort to the manual approach. Usually, not being able to reach the needed certainty for the case was only realized after extensive verification of many cell-level decisions.

We sought to remedy this by finding another intermediary level, that had more automatic support than the one above, but less than only getting a final percentage. To find opportunities for automatic support we analyzed the bias-based error in our underlying model. We found that most errors are somewhat systematic; visually similar patches might all be assigned the "wrong" classification. For instance, the threshold for brown staining intensity to be considered positive may differ between cases.

Based on this, we added an algorithm for unsupervised visual similarity clustering to our system and sought to design the interaction such that the user could work by only making decisions on a cluster level. The user interface for this mode of interaction is depicted in 6.

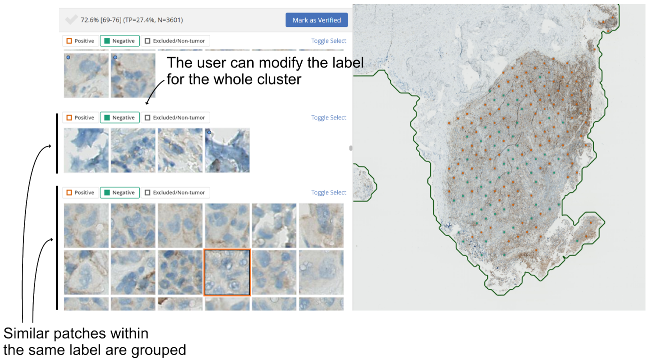

72.6% [69-76] (TP=27.4%, N=3601)
Mark as Verified
Positive
Negative
Excluded/Non-tumor
Toggle Select
The user can modify the label for the whole cluster
Similar patches within the same label are grouped

Figure 6: Patches with the same predicted label are grouped by clusters in the left-most panel. In this example, the cluster with four patches should be reclassified as non-tumor. An experienced pathologist is able to do this just by looking at the group of patches.

In this prototype, the user can choose to look at the resulting percentage (e.g. `72.6% [69-76] N=3601`), or to view the first few patches of each cluster, or to expand clusters and inspect their constituent patches. Additionally, clusters are ordered by uncertainty, and patches within the cluster are also ordered by uncertainty. The intent is that the user hopefully can detect errors in only the first few clusters and then accept the rest.

A typical, multi-level workflow when using this would be as follows:

(1) Open the case and initiate the use of the tool
(2) (*top level*) Review the overall assigned percentage. Is it reasonable given the overall look of the tissue? If the confidence interval is far from a treatment cut-off, accept the result. Otherwise continue.
(3) (*individual corrections*) Is the grid dense? If not, start reviewing and correcting the patches of the top-most clusters. Observe the updated percentage and the confidence. Stop when you're making fewer corrections per cluster.
(4) (*batch correction*) If the grid is dense, and there are over 500 patches, start reviewing the top-most clusters; based on its first patch, does it have the correct classification? If not, correct the classification for the entire cluster. Observe the updated percentage and the confidence.
(5) (*individual correction*) Check the patches in the cluster; does any patch "stand out" as not belonging to the cluster? Correct the patches by dragging them to the correct category, they will automatically be assigned another cluster of that type.
(6) (*batch correction*) Proceed through a few clusters, once no or few errors are detected, the rest is probably correct.

Evaluation

We presented this multi-level version of the tool to three pathologists that had not been part of the design process in a small qualitative assessment. The three pathologists were presented the tool for the first time. We wanted to know whether the prototype could be clinically useful and more specifically, whether it seemed the pathologists could learn multi-level strategies that allowed them to balance detailed control, spent time and diagnostic quality. Our goal was primarily to assess the concept's viability for further empirical efforts.

We found a recurring theme of initially wanting to drill-down to cell level. Pathologists reported that they would need some "alone time" to learn what kind of systematic errors the prediction was making, and correlate this to the overall appearance of the case. When asked whether they thought they would be able to learn when to work at which level of detail, they were tentatively positive, but stating that time would tell for certain.

To us, it seemed the design had potential in allowing them to work with sometimes inaccurate models, but also, by moving between levels. Through drill-down we hope that they might learn to calibrate their trust towards working at the right level as appropriate. It could be that a more global, model-level understanding can be achieved by interacting with local justifications like ours, over time. By contrast, user interfaces where human-ML collaboration becomes a verification cliff does not as readily afford this, as the manual approach and the assisted are completely disjunct.

While the results from such a small user study are mostly anecdotal at this point, we are planning to evaluate this aspect more extensively in future research.

4 DISCUSSION

While our concept of verification staircases is early work, we believe it has connections to many of the same issues that research on explainable and transparent intelligent tools seek to address.

For instance, many of the principles outlined for Explanatory Debugging [3] are imbued in our concept. Such as: being iterative, being sound & complete, not overwhelming and being actionable. The major difference is that our proposed explanations do not correlate predictions to the inner workings of the model, but instead to the underlying phenomena viewed at different fidelities.

The need for enabling user feedback for explanations [12] is facilitated by excluding references to inner workings of the model, letting the images of the domain problem always act as the shared language to create common ground for communication. It is noteworthy that this interaction affords continuous learning of the machine learning component by enabling the corrections to become training data for future iterations [4].

Enabling global model understanding through repeated exposure with local justifications is similar to the strategy employed by the LIME technique [9].

Our current design aids the user in detecting errors, e.g., by sorting patches and clusters on confidence. We then rely on that the user will be able to learn which end of the model's accuracy distribution they are in, or at least, the suitable amount of validation effort to spend. There exist other approaches to facilitating error detection and determining the accuracy of classifiers [2] that could be interesting to incorporate in future versions.

A limitation of our current prototype is that a user's correction of single patches or clusters affect only the directly involved patches, clusters and the overall ratio. We have experimented with versions where the model is fine-tuned using this input and the predictive output is updated, in an interactive machine learning manner. However, this kind of global updates creates a lack of control for which we are yet to find good interaction design solutions that suit our safety-critical domain. We believe this is an interesting area of future research.

ACKNOWLEDGMENTS

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP)

REFERENCES

[1] Gilbert Cockton. 2006. Designing Worth is Worth Designing. In *Proceedings of the 4th Nordic Conference on Human-computer Interaction: Changing Roles (NordiCHI '06)*. ACM, New York, NY, USA, 165–174. https://doi.org/10.1145/1182475.1182493 event-place: Oslo, Norway.

[2] Alex Groce, Todd Kulesza, Chaoqiang Zhang, Shalini Shamasunder, Margaret Burnett, Weng-Keen Wong, Simone Stumpf, Shubhomoy Das, Amber Shinsel, Forrest Bice, and Kevin McIntosh. 2014. You Are the Only Possible Oracle: Effective Test Selection for End Users of Interactive Machine Learning Systems. *IEEE Transactions on Software Engineering* 40, 3 (March 2014), 307–323. https://doi.org/10.1109/TSE.2013.59

[3] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15)*. ACM, New York, NY, USA, 126–137. https://doi.org/10.1145/2678025.2701399 event-place: Atlanta, Georgia, USA.

[4] Martin Lindvall, Jesper Molin, and Jonas Löwgren. 2018. From Machine Learning to Machine Teaching: The Importance of UX. *Interactions* 25, 6 (Oct. 2018), 52–57. https://doi.org/10.1145/3282860

[5] Shane T. Mueller, Robert R. Hoffman, William Clancey, Abigail Emrey, and Gary Klein. 2019. Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI. (Feb. 2019). https://arxiv.org/abs/1902.01876v1

[6] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. 2019. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. *arXiv:1909.12475 [cs, stat]* (Nov. 2019). http://arxiv.org/abs/1909.12475 arXiv: 1909.12475.

[7] Joaquin Quiñonero-Candela (Ed.). 2009. *Dataset shift in machine learning*. MIT Press, Cambridge, Mass.

[8] Maithra Raghu, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. 2019. The Algorithmic Automation Problem: Prediction, Triage, and Human Effort. *arXiv:1903.12220 [cs]* (March 2019). http://arxiv.org/abs/1903.12220 arXiv: 1903.12220.

[9] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778 event-place: San Francisco, California, USA.

[10] Thomas B. Sheridan. 2018. Comments on "Issues in Human–Automation Interaction Modeling: Presumptive Aspects of Frameworks of Types and Levels of Automation" by David B. Kaber. *Journal of Cognitive Engineering and Decision Making* 12, 1 (March 2018), 25–28. https://doi.org/10.1177/1555343417724964

[11] Thomas B. Sheridan and William L. Verplank. 1978. Human and Computer Control of Undersea Teleoperators. https://doi.org/10.21236/ada057655

[12] Alison Smith and James J Nolan. 2018. The Problem of Explanations without User Feedback. (2018). Position paper presented at the IUI'18 Workshop on Explainable Smart Systems.

[13] Antonio Torralba and Alexei A. Efros. 2011. Unbiased look at dataset bias. In *CVPR 2011*. 1521–1528. https://doi.org/10.1109/CVPR.2011.5995347 ISSN: 1063-6919.

[14] K. J. Vicente and J. Rasmussen. 1992. Ecological interface design: theoretical foundations. *IEEE Transactions on Systems, Man, and Cybernetics* 22, 4 (July 1992), 589–606. https://doi.org/10.1109/21.156574