1 Answers to the comments of the Associate Editor

1. **Editor:** *The reviewers generally liked the goal of the paper, but there were some concerns*

   Answer: Thank you for your comments on our paper. We went through the manuscript carefully and made changes, following the comments. We hope that these changes address the concerns.

2. **Editor:** *The evidence of potential impact relies on a statement of interest and use by 3 anonymous companies. As the special issue is designed to be a showcase of what ML is doing for society, it is problematic for the special issue to have the impact be completely anonymous. As outlined in our CFP, we would particularly like expert commentary, and possibly industrial co-authors.*

   Answer: That makes sense. We have now added the names of the three companies that have adopted or were influenced by our "Beat the Machine" system (namely, Integral Ad Science, oDesk, and Tagasauris). In a sense, the whole paper is "expert commentary" — these problems arose and were solved largely while the authors were working for the companies in question. As we now clarify in the article (please see section for specifics), all three authors are industrial co-authors (despite the academic affiliations). We realize that this was requested in the call for papers; apologies for not clarifying in the original submission.

3. **Editor:** *The authors have a conference paper (archival) that has quite a lot of overlap with the submitted paper. As far as we can tell the main contribution beyond the conference paper is Section 6: Impact in Industrial Deployments, which is the section that is not substantial enough to allow the paper to fit our CFP at the moment, as discussed above. Since the MLJ paper should have at least 30% more material than in the conference version, that section would need to be expanded to be more substantial to fulfill the requirement. The paper should also address how the current work differs from the conference work.*

   Answer: We would like to clarify that the prior paper is not an archival conference paper, it is a paper at a specialized workshop (namely HCOMP11 at KDD). The current submission adds to the prior work. For example, it introduces a proper formalization of the concept of unknown unknowns. It also discusses issues of deployment, that were not included in the former.

   [Foster here: it has been the policy of the MLJ to differentiate in prior work between top-level archival conference papers and specialized (not stringently reviewed) workshop papers, and to give much more leeway to the republication of results from the latter. This is important to the field in at least two intimately related ways. First, it fosters the quick dissemination and discussion of preliminary results, which is generally good for a fast-moving field. Second, it promotes the later rigorous review of those results, in the context of journal submission.]

4. **Editor:** *Some of the reviewers did not see how the information from the Beat the Machine system was used in improving the underlying ML classifier.*

   Answer: Indeed. This is not a paper about the machine learning algorithms themselves, but about another stage in the data science process: the evaluation of machine learning algorithms (and the resultant models). In our view, proper evaluation is a very important topic for machine learning research. Using the discovered unknown unknowns to improve the ML classifier is an interesting direction for future research, but not the focus of this paper.

5. **Editor:** *Steps in data preparation and machine learning seem to be missing.*

   Answer: We now provide the details about the classifiers, how they were trained, and provide other details requested by the reviewers.

6. **Editor:** *We encourage the authors to submit a revision to address the problems discussed above, in addition to the other issues mentioned by the reviewers. Please prepare a point by point response to the reviewers when the paper is resubmitted. The due date for the resubmission is February 20.*

   Answer: n/a

2 Answers to the comments of Reviewer 1

1. **Reviewer 1:** *In this paper the authors propose a system called Beat the Machine. Given a classifier, Beat The System solicits users to submit examples. Users are given reward for submitting examples that are classified incorrectly by the system and with high confidence. The authors propose a few iterations of the system design which mainly differ in the reward given for a submission. Finally, the authors present experimental results on a hate speech detector and an adult content detector.*

   Answer: n/a

2. **Reviewer 1:** *I have the following major concerns about this paper:*

   Answer: n/a

3. **Reviewer 1:** *1. Rather than being an application of machine learning, this paper is more about finding the limitations of a machine learning system using human computation to identify the so-called "unknown unknowns".*

   Answer: The reviewer is correct. Perhaps this is a topic for debate, but based on our (comparatively quite long) experience in applying machine learning techniques to real problems, "finding the limitations of a machine learning system" is a very important aspect of doing machine learning. We need to understand how the classifier is expected to behave when released in production. Very often in practice, the data on which machine learning algorithms are trained is actually not a representative sample of the data to which they will be applied, because getting the latter often is prohibitively expensive. Even when one can sample from the target distribution, often there are very important small disjuncts that are missed. The BTM approach offers a proactive evaluation technique that tries to identify problems in learned models before they become real business problems for the stakeholders.

4. **Reviewer 1:** *2. The paper gives experimental evidence to suggest that beat the machine system finds "regions of the problem space" where a classifier doesn't know that it is performing bad there. However, there is no evidence of impact or potential impact on science or society other than the claims of anonymous companies showing interest in the proposed system.*

   Answer: Yes, sorry about that. (Please see response to the editors above.) We have now identified the companies in the paper.

5. **Reviewer 1:** *3. Many important details are missing in data preparations and experiment setup. For example, here are some of the questions I had about the experiments:*

   Answer: see next

6. **Reviewer 1:** *What was the classifier that was challenged in either case? What was the training data that was used for this purpose? How many training examples were there for training?*

   Answer: Revised. For both tasks (adult and hate speech), the underlying classifier was a model trained using logistic regression. Each classifier was trained with 20,000 URLs. Featurization is described briefly in the revision.

7. **Reviewer 1:** *How many humans participated in this study?*

   Answer: Revised. For the hate speech task, a total of 28 workers participated, we collected 500 URLs. For the adult content task, a total of 25 workers participated and again we collected 500 URLs. We added the clarification in the paper.

8. **Reviewer 1:** *What guidelines did they use for submitting URLs?*

   Answer: In Figure 3, we show the interface shown to the the crowdsourced users, together with the guidelines. The instructions were purposefully brief and we did not give any significant background to the users about the classifier. The goal was for the users to treat the classifier as an opaque black box, that can only be probed for vulnerabilities, and would return only the classification result for a URL.

9. **Reviewer 1:** *There is a not adequate detail on stratified random examination. How were the "random" URLs selected? Were they completely selected at random or were they some random URLs that contained hate/racist/adult content words? If they were just random URLs with no relevance to the problem at hand, I am not even sure it is a fair comparison considering that the space of URLs is huge.*

   Answer: We selected the URLs from the stream of web pages that are classified daily by the classifier. Each URL had an expected misclassification cost assigned by the classifier. (When the classifier was uncertain about the classification, the expected misclassification cost was high, and vice versa.) We then formed k equal-width bins, splitting the URLs based on their expected misclassification cost, and we picked an equal number of URLs from each bin: Out of the total of N URLs used for testing, we had $\frac{N}{k}$ URLs from each bin.

   The "sampling using keywords" idea is insightful. (Please see the KDD-2010 paper by Attenberg & Provost for a closely related discussion) Generally, since the existing models are word-based models, the stratified examination does a sophisticated version of this. However, the URLs returned by keyword-based sampling tend to miss the exact regions that the current paper focuses on: those that one hasn't thought of previously in the sampling of the training data. (Generally, as described by A&P 2010, the firm's models would have been trained in the first place on keyword-sampled pages.)

10. **Reviewer 1:** *4. It was disappointing to see that the paper stops after identifying "unknown unknowns". This paper would have had significant impact if the information found from Beat The Machine system was used in improving the underlying machine learning classifier. In this sense, the realized or potential impact of the content presented in this paper is much lower than what could have been achieved.*

    Answer: We share the reviewer's opinion that there is substantial impact to be had if one were to design methods to improve learning with the unknown unknowns. That is indeed not what this paper is about. Hopefully this paper can have significant impact in the research world, by challenging researchers to do so.

    The focus of this paper is on the actual impact that these methods have had: improving and accelerating the evaluation of learned classification systems, specifically with respect to identifying unexpected areas of failure, and more generally changing the way firms are thinking about evaluating their systems. If this leads researchers to then focus on this area and invent great ways to improve systems in response, then this work would have had exactly the sort of impact many of us have desired of applications-oriented papers—feeding back to influence the topics of more academic research (see e.g. the Provost & Kohavi MLJ editorial from 1998). We can hope.

11. **Reviewer 1:** *Please summarize the paper's claims about impact achieved from a machine learning advance: The main practical impact of this paper is a system that results in a new approach for testing and debugging automatic machine learning models.*

    Answer: n/a

12. **Reviewer 1:** *Impact: In my opinion the significance of this impact is likely to be minor. The paper gives experimental evidence to suggest that beat the machine system finds "interesting areas of space" where a classifier doesn't know that it is performing bad there. However, there is no evidence of impact or potential impact other than the claims of anonymous companies showing interest in the proposed system. There are no details of the impact made by the proposed system in any of these companies.*

    Answer: We have now added the names of the companies in the paper. We have described how the companies have implemented or are implementing the systems and have changed their processes for evaluating systems, which is substantial real-world impact.

13. **Reviewer 1:** *Novelty: This paper is about challenging humans to find examples where a classifier is likely to be wrong with high confidence. While there is no new machine learning technology developed in this paper, the idea of using humans to find challenging examples for a machine learning system does seem novel.*

    Answer: n/a

14. **Reviewer 1:** *Problem description: This paper first describes a system and is subsequently applied to a problem domain. In particular, the developed system is applied to a hate content detector and an adult content detector.*

    Answer: n/a

15. **Reviewer 1:** *Data preparation: Many steps of data preparation are missing. The authors do not reveal any details about the original classifiers or the data that was used for training the system. They do not give any details on how the humans went about finding the URLs. Please see above for a partial list of unanswered questions.*

    Answer: We now provide additional details about the underlying classifiers, and give the instructions given to humans that undertook the BTM task. We hope that the provided answers satisfy the raised concerns.

16. **Reviewer 1:** *Machine learning: The machine learning component of the paper is limited. The paper is mainly about building a system where humans can submit examples in a game like environment to challenge a machine learning system.*

    Answer: Please see responses above.

17. **Reviewer 1:** *Results: The methodology seems appropriate for the system described. However, the system itself stops at identifying unknown unknowns and does not do anything interesting with what is identified.*

    Answer: We disagree with the sentiment of the reviewer here. We believe that identifying the unknown unknowns, and warning the stakeholders of the learned system about the *unexpected* weaknesses, is extremely interesting, important, and in fact crucial for robust operation (in such environments) of a learned model in production. The fact that three companies have invested in implementing/incorporating such systems provides supporting evidence. We are not aware of any other system that can achieve this goal.

18. **Reviewer 1:** *Domain expert: The authors do provide comments from company founders and data scientists about this system being deployed in their companies However, the names of these companies and the names of the persons who spoke for their behalf are not revealed. Nor is the actual impact in those companies quantitatively revealed.*

    Answer: We provide now the names of the companies, and describe how the paper is essentially a collaboration of industrial coauthors (who therefore speak on behalf of the companies). We discuss how the companies have changed their procedures based on this work. We show in the context of an actual classified data stream from Integral Ad Science that the procedure indeed does find unknown unknowns. If these go undetected, Integral faces increased reputation risk.

    The reviewer is correct that we do not provide quantitative results (for example) on whether the firms have actually avoided catastrophes using the system. That is a an extremely interesting angle for which we do not currently have quantification. (For example, Integral Ad Science would like to avoid mistakenly putting an ad on a hate speech page followed by a catastrophic media event for its client. Quantifying the counterfactual that you have avoided a catastrophic event is quite difficult.)

    We would argue that even if the companies have implemented the systems purely as a sort of insurance policy, the fact that they have shows substantial impact of the work. (I'm happy with my life and car insurance policies if they never pay off quantitatively.)

19. **Reviewer 1:** *Infusion: The authors claim that many (unidentified) companies are currently using/ or interested in using the proposed system.*

    Answer: please see above.

20. **Reviewer 1:** *Lessons: The paper does identify a potentially interesting problem to the machine learning community.*

    Answer: n/a

3 Answers to the comments of Reviewer 3

1. **Reviewer 3:** *This paper describes the impact of a technique (Beat the Machine) for making use of crowd sourcing in order to identify test examples that are: (a) likely to be classified incorrectly by a learned classifier, (b) about which the classifier is confident, and (c) that, if classified incorrectly, can be extremely costly. These types of examples arise in learning scenarios where there is significant class imbalance and where misclassifying an instance from the negative class is expensive either monetarily or otherwise.*

   Answer: n/a

2. **Reviewer 3:** *Specifically, this technique, according to the authors, has (a) changed the way one medium-scale company evaluates its systems. This company does "massive-scale webpage classification in the online advertising space." This same company has invested the industrial development of Beat the Machine and is applying it to tasks beyond the initially intended application. (b) directly influenced the workflow design of a large firm that runs an online labor marketplace. For instance, Beat the Machine has been deployed to test a job classification engine. (c) has been deployed as part of an image-tagging service for a third company.*

   Answer: n/a

3. **Reviewer 3:** *More generally, Beat the Machine provides a new approach for testing machine learning models where the misclassification of rare, but systematic, cases can be problematic, if not catastrophic.*

   Answer: n/a

4. **Reviewer 3:** *2. Impact: The companies' names are not given, nor is the impact quantified. Here we are relying entirely on the honesty of the authors. Still, I judge the impact to be major. The authors make a good case for the significant negative consequences of *not* finding the types of examples identified through Beat the Machine. That at least three companies have adopted it for different types of applications attests to its utility.*

   Answer: The revision states the companies as well.

5. **Reviewer 3:** *3. Novelty: The paper does not describe an application of ML per se. Instead, it addresses a type of domain that is difficult for classifier-learning systems. It provides an approach for gathering test examples that are selected precisely for their ability to challenge the learned classifier. This notion is similar to employing experts to attempt to breach security mechanisms, in order to ensure their strength. To the best of my knowledge, the authors' approach is novel.*

   Answer: n/a

6. **Reviewer 3:** *4. Problem description: In general, the problem domain is any domain with the properties described above. The problem domain used as an example throughout the paper is classification of web pages as containing/not containing objectionable content such as "hate speech." The problem domain is described sufficiently to be understandable.*

   Answer: n/a

7. **Reviewer 3:** *Unfortunately, in its formal definition of "unknown unknowns", the paper conflates probability with cost, is unclear on the use of "expected valued", and is highly repetitive. More specific details on these points can be found below.*

   Answer: Indeed - we shared the concern about the mix of probabilities and expected cost. As shown in Equations 1 and 1, the expected misclassification cost is directly related to the returned probability estimate across the potential labels for each example, hence we often use the terms interchangeably when we do not expect a confusion. We opted to use the term "expected misclassification cost" due to cases where the prediction maybe highly uncertain but still have low expected misclassification cost (and therefore be inconsequential). Consider for example the case of adult content classification into four categories: "G" for general-audience, "P" for content requiring parental supervision, "R" for adults-only, and "X" for hard-core porn. The misclassification cost of $X \to G$ is typically high, while the cost for a

$P \rightarrow R$ is typically much lower. If we simply rely on probability values, an example classified as 50% X and 50% G is equally uncertain to an example classified as 50% P and 50% R. However, the expected misclassification cost of the former is much higher, showing the need to introduce costs in the definition of "unknown unknowns." We've added some clarification in the text; we are happy to add more if you feel that that is inadequate.

8. **Reviewer 3:** *5. Data preparation: The mechanism for gathering data through Beat the Machine is clear.*

   Answer: n/a

9. **Reviewer 3:** *Though the focus of the paper is Beat the Machine (BTM) and identification of web pages with objectionable content is only one example where it could be used, this particular domain *is* the one on which the BTM approach is validated. It is clear that the examples are web pages, but details are not given on how they are featurized for learning. For example, footnote 1 on page 3 says web pages are represented by their words, links, images, metadata, etc. Are images really represented in the features? How? (Because this work was done for a company that is not even named, I imagine the data are proprietary.)*

   Answer: We apologize for the confusion—and for a mistake that we made. The general answer is that we use as subset of the featurization used by the company. We had included a generic description based on what they do. For (some) images, the company uses a third-party classification service that returns information about the content of the image (e.g., if there is suspicion for nudity). We have clarified the featurization that we used, and indeed the featurization does not actually include the image information, but is based on the words on the page, metadata, title, etc.

10. **Reviewer 3:** *6. Machine learning: As for the previous question, the point of the paper is not any specific machine learning system or algorithm. But one is used to validate the approach. However no details at all are provided for the algorithm, so the specific experiments detailed in Section 5 ("Experimental Studies") would not be reproducible. Still, the experiments are described in enough detail that we can understand what was done and how.*

    Answer: We have added details about the underlying classification system, e.g., size of the training set, algorithm used for training the classifier. (Given that BTM treats the underlying classification model as a black box, in essence it can use any induction algorithm; we used logistic regression.) The reviewer is correct about reproducibility—very specifically: we cannot share the exact training data, so the exact results could not be reproduced. However, we hope that they are replicable in a broader (and probably more interesting) sense: if someone else were to build a hate-speech or adult-content classifier using standard ML procedures, sample a large set of new web pages, and then run a BtM system, that they would get similar results.

11. **Reviewer 3:** *7. Results: The experiments are carried out on a specific domain (web page objectionable content of two types: hate speech and adult content). BTM is compared to stratified random examination. Basically, the authors are showing that for a specific model builder for a specific domain (or pair of domains), it makes more sense to use BTM than the method currently used for that system. The results are compelling, and the questions asked are appropriate.*

    Answer: n/a

12. **Reviewer 3:** *To the extent that we care that BTM is making an impact in a real application, this is a fine evaluation. It does not prove its utility more generally.*

    Answer: We agree that there could possibly be a lot of followup work on BtM.

13. **Reviewer 3:** *8. Domain expert: Given the adoption of the BTM technique by at least three companies, we can take this as confirmation by domain experts of BTM's utility. The paper is written in such a way that it would be understandable by an expert with data that had the general properties of concern to the authors. Most of the results would be interpretable by domain experts who were familiar with classifier learning.*

    Answer: n/a

14. **Reviewer 3:** *9. Infusion: The authors clearly describe how BTM has been incorporated into deployed systems.*

    Answer: n/a

15. **Reviewer 3:** *10. Lessons: One of the strengths of the paper is that it details the evolution of BTM, explaining clearly why earlier versions of it did not achieve the authors' goals as well as the final version.*

    Answer: n/a

16. **Reviewer 3:** *Though I have recommended accepting this paper with minor revisions, by "minor" I am referring to overall content and experiments. The writing itself could use more significant work. Because I am sympathetic to the goals of the paper and would be happy to see it accepted, below I provide more detail on the changes I would recommend.*

    Answer: Thanks.

17. **Reviewer 3:** *Issues of clarity: Page 2, lines 28-35. Unknown unknowns are introduced as being examples on which a classifier is confident but actually wrong. In the final two lines of the paragraph, an unknown unknown is suddenly described in terms of misclassification cost. Throughout the paper, probability, confidence, and cost are confounded. I would recommend defining unknown unknowns in terms of confidence/probabilities, and keeping cost out of that definition. Cost can then be introduced on top of the definition, as unknown uknowns are, for a non-trivial number of domains, precisely those cases whose misclassification can be costly. The fundamental problem addressed by the paper is clear and important, but the attempt at a formal definition is problematic as is.*

    Answer: Please see the answer above (item 7) for the justification of our choice to use misclassification cost in the definition, instead of probability and confidence.

18. **Reviewer 3:** *Page 3, lines 23-26: again, cost and probability are muddled here.*

    Answer: Fixed.

19. **Reviewer 3:** *Section 3 is especially problematic. First, a direct relationship is assumed between confidence and cost, and that isn't necessarily true. As above, I recommend defining unknown unknowns (and the other types of known/unknowns) in terms of a system's level of confidence and its probability of being incorrect. (For Fig 1, for example, the y axis would be the system's estimated confidence. The x axis would by the system's actual probability of being incorrect.)*

    Answer: Please see the answer above (item 7) for the justification of our choice to use misclassification cost in the definition, instead of probability and confidence.

20. **Reviewer 3:** *Definition 1: Be careful about parallel sentence construction. "We denote by...., and ... be...." should be "Let .... be ..., and let .... be...."*

    Answer: Fixed.

21. **Reviewer 3:** *This definition says that the *actual* misclassification cost ExpCost(x)... Using expected cost notation for the actual cost, which is a constant for any given x, is problematic.*

    Answer: Agreed. We removed the hat notation for the estimated cost, calling it just *ExpCost*, and we use the term *Cost* for the actual misclassification cost.

22. **Reviewer 3:** *"prediction-time classification"? This is not a term I've heard; nor could I find it. Rewrite or delete the sentence that contains it.*

    Answer: Removed the term "prediction-time".

23. **Reviewer 3:** *In the next paragraph, you say "this model is likely to encounter examples eliciting a high degree of predicted uncertainty". I don't see why this is the case. Please clarify.*

    Answer: Fixed.

24. **Reviewer 3:** *Page 6, around line 39/40, you say "data is gathered by some random process, for instance via active learning". While active learning can employ a random process, this is not generally the case. What are you really trying to say here?*

    Answer: We removed the term "random" that was causing confusion.

25. **Reviewer 3:** *The paragraph that begins "Finally, from Figure 1, we can see" looks textually like it's still part of the description of Definition 2, but it clearly isn't. The formatting needs to be fixed.*

    Answer: Fixed.

26. **Reviewer 3:** *In addition to the probability/cost conflation, this section is highly repetitive. The paper takes 6 pages to set up the problem and then 6 to detail the system, validate it, and give evidence of its impact. The former can easily be condensed. In fact, condensing it would likely make it more clear.*

    Answer: We appreciate the comment. The paper is indeed rather verbose early on, because many readers of earlier versions of the draft indicated that they were confused about the concept of unknown uknowns. We tried to add additional (potentially reduntant) material, mainly for reasons of readability. We understand that we may disappoint readers that want to get quicker to get main contribution, but we believe that this is a balance that we need to strike.

27. **Reviewer 3:** *Issues with Section 5, Experimental Studies: In the comparison with stratified random examination, you say that "they are designed to assess different quantities". So make it clear up front that you expect them to be different and that you do the experiment to verify the parts of the example space that each one focuses on.*

    Answer: We rephrased, trying to make clear that we expect the two techniques to focus on different parts of the space.

28. **Reviewer 3:** *You refer to "natural error rate". What do you mean by this?*

    Answer: Rephrased to clarify that this is the error rate when tested with random URLs.

29. **Reviewer 3:** *In the comparison of error severity, you say "1000 means that the classifier was certain of one class and the actual class was the other." Does this hold true for both majority→minority and minority→majority? Or is it only in cases where the classifier thought "majority" and it was really "minority"?*

    Answer: In this case, we refer only to cases where the classifier thought that page contained no offensive content, which is the final design of the BTM system. (See Section 4 for the rationale for this choice.)

30. **Reviewer 3:** *Again, in lines 39-41 on page 10, there is a confounding of cost and probability.*

    Answer: Fixed.

31. **Reviewer 3:** *The final paragraph of the same subsection is grammatically problematic and awkward. It needs to be fixed.*

    Answer: Fixed

32. **Reviewer 3:** *Beat the Machine is sometimes written with quotes, sometimes without, sometimes italicized, sometimes not, sometimes upper case, sometimes not... In general, this needs to be cleaned up for consistency.*

    Answer: We have unified the different styles used to refer to the BTM system.

33. **Reviewer 3:** *Similary, unknown unknowns is written in all sorts of formats in all sorts of places, even within the same paragraph. This can be very distracting to the reader. The same is true for known unknowns and all other variations.*

    Answer: Fixed.

34. **Reviewer 3:** *Abstract "predictive-model-based" → "predictive model-based"*

    Answer: Fixed.

35. **Reviewer 3:** *“cases that do not reveal” → “cases that may not reveal”*

    Answer: Fixed.

36. **Reviewer 3:** *Section 1 “learing” → “learning”*

    Answer: Fixed.

37. **Reviewer 3:** *“based on models... and produces” → change “produces” to “produce”*

    Answer: Fixed.

38. **Reviewer 3:** *“performance in unseen data” → “performance on unseen data”*

    Answer: Fixed.

39. **Reviewer 3:** *“ML research” → “machine learning (ML) research” After that, can use ML.*

    Answer: Fixed.

40. **Reviewer 3:** *page 2, line 37/38: “AUC” used before defined.*

    Answer: Fixed.

41. **Reviewer 3:** *Section 2 “to prepare to deal with” → “to prepare for”*

    Answer: Fixed.

42. **Reviewer 3:** *One of the requirements for crowdsourcing is having a problem that the “average person” is able to address. The “hate speech” domain is one such problem, and there are others as well, some of which are described in Section 6. I recommend giving several such examples as early as Section 2, so that the reader is clear that BTM can indeed have broad applicability.*

    Answer: We introduced a forward pointer to Section 6, to indicate that the system has been used in a variety of different settings.

43. **Reviewer 3:** *“generaliry” → “generality”*

    Answer: Fixed.

44. **Reviewer 3:** *“it can be very costly to find very few positive examples” - awkward. rephrase.*

    Answer: Fixed.

45. **Reviewer 3:** *“far less that” → “far less than”*

    Answer: Fixed.

46. **Reviewer 3:** *Page 3, lines 47-49: The transition to active learning is appropriate, but awkwardly written.*

    Answer: Fixed.

47. **Reviewer 3:** *“where we would think to find errors” → “where we would expect to find errors”*

    Answer: Fixed.

48. **Reviewer 3:** *“a system to use human workers” → “a system that uses human workers”*

    Answer: Fixed.

49. **Reviewer 3:** *“confident and wrong” → “confident but wrong”*

    Answer: Fixed.

50. **Reviewer 3:** *“workers that discover” → “workers who discover”*

    Answer: Fixed.

51. **Reviewer 3:** *“participation in the tasks” - be more specific about “the tasks”*

    Answer: Fixed.

52. **Reviewer 3:** *"We describe our first experiences by the live deployment..." Huh? Awkward as written.*

    Answer: Fixed.

53. **Reviewer 3:** *Section 3 "The task of a classification is to construct" → "The learning task is to construct"*

    Answer: Fixed

54. **Reviewer 3:** *"gives birth to" → "gives rise to"*

    Answer: Fixed.

55. **Reviewer 3:** *"that your model is known" → "that a model is known"*

    Answer: Fixed.

56. **Reviewer 3:** *"discussed previously" → "discussed above"*

    Answer: Fixed.

57. **Reviewer 3:** *Caption for Figure 2: refers to the figures as "top" and "bottom", but they're side by side.*

    Answer: Fixed.

58. **Reviewer 3:** *"Call such examples" → "We call such examples"*

    Answer: Fixed.

59. **Reviewer 3:** *"examples that the model is quite certain a correct label can be assigned" - incorrect grammar*

    Answer: Fixed.

60. **Reviewer 3:** *"we see an a region"*

    Answer: Fixed.

61. **Reviewer 3:** *Section 4 "can be challenge" → "can be a challenge"*

    Answer: Fixed.

62. **Reviewer 3:** *"errors would be misclassified" → "errors will be misclassified" (for verb tense consistency"*

    Answer: Fixed.

63. **Reviewer 3:** *"examples truly of" → "examples of"*

    Answer: Fixed.

64. **Reviewer 3:** *"This is problematic" - be careful to say what "this" is.*

    Answer: Fixed.

65. **Reviewer 3:** *"examples truly in the minority class" → examples whose true class is the minority class.*

    Answer: Fixed.

66. **Reviewer 3:** *"relatively accurate classifier, with 95% error rate" - surely you don't mean this.*

    Answer: :-) Fixed.

67. **Reviewer 3:** *even "outlier" cases can cause significatn damage" - be more precise about what you mean by "outlier"*

    Answer: Fixed.

68. **Reviewer 3:** *"client's expectations" → "clients' expectations"*

    Answer: We refer to a single client, whose expectations were not met.

69. **Reviewer 3:** *"hackers that are hired" → "hackers who are hired"*

    Answer: Fixed

70. **Reviewer 3:** *"client's expectations" → "the client's expectations"*

    Answer: Fixed

71. **Reviewer 3:** *Section 4.1 The first sentence is internally repetitive and awkward.*

    Answer: Fixed.

72. **Reviewer 3:** *There are many issues in this and subsequent sections with verb tense. I suggest describing the different versions of BTM in the present tense. But more importantly, there needs to be a check for consistency.*

    Answer: Fixed

73. **Reviewer 3:** *Also, "Design 1" is labeled as such, but subsequent designs are titled only by what they added to the previous. Clearly mark Design 1, Design 2, Design 3 as such.*

    Answer: Fixed

74. **Reviewer 3:** *Fig 3 did not print clearly for me. If included, it needs to be a higher- resolution screen shot.*

    Answer: We increased the size and resolution of the image.

75. **Reviewer 3:** *Page 8, lines 33-34. So what exactly did this mean on a practical level? That they had few people choose to do the task? That few would return to do it?*

    Answer: This design leads to a short user engagement, and high rates of abandonment. We added this clarification in the paper.

76. **Reviewer 3:** *Page 8, lines 47-52. Get rid of the parentheses. They're unnecessary and unhelpful.*

    Answer: Fixed

77. **Reviewer 3:** *"misclassification cost are given a the reward is small" - huh?*

    Answer: Fixed.

78. **Reviewer 3:** *Section 5 "we described the concept of the" → "we defined"*

    Answer: Fixed.

79. **Reviewer 3:** *"a gamified structure" - "gamified"???*

    Answer: We think that the term "gamified structure" describes accurately the BTM system. We can remove the term, if you feel strongly otherwise.

80. **Reviewer 3:** *"with the configuration details" → "with the final configuration details". Basically, be clear *which* configuration you're using. You described several.*

    Answer: Fixed.

81. **Reviewer 3:** *"In the application domain, the standard procedure..." Be clear whether you're really talking about a specific application domain or a more general class of applications.*

    Answer: Fixed.

82. **Reviewer 3:** *Figure 4 caption: First sentence is awkward/unclear. Change "mistake" to "error" throughout.*

    Answer: Fixed.

83. **Reviewer 3:** *“However, you may have noted that” → “Note that”*

    Answer: Fixed.

84. **Reviewer 3:** *“Figure s4(a)” → “Figures 4(a)”*

    Answer: Fixed.

85. **Reviewer 3:** *“modeling mistakes” → “errors”*

    Answer: Fixed.

86. **Reviewer 3:** *References Check the Weiss reference. I checked Weiss’s publications page, and the venue of publication looks wrong to me.*

    Answer: Fixed.

4 Answers to the comments of Reviewer 4

1. **Reviewer 4:** *Please summarize the paper’s claims about impact achieved from a machine learning advance. Rather than machine learning advance, this paper is about human computing. In particular, a Beat the Machine game is designed for humans to provide examples on which the ML algorithm doesn’t know it was wrong. This can be viewed as a new variant of previous game-based human computing efforts.*

    Answer: n/a

2. **Reviewer 4:** *Impact: This is a nice idea that, according to the paper, has been adopted by three companies and therefore has had real impacts.*

    Answer: n/a

3. **Reviewer 4:** *Novelty: This is not really ML but human computing designed to find weaknesses of a classifier. I have not seen this particular idea before.*

    Answer: In a strict sense we agree – if we are considering ML to focus just on algorithms for doing the ML itself, and not the greater process. The greater process does seem to be within the scope of the special issue.

    Another perspective is that BtM is a new sort of combined human/computer discover, which at one time was considered to be Machine Learning. We have not waxed philosophical about that, because to us it seems that the paper is relevant enough, and that such a discussion would detract from the focus of the paper.

4. **Reviewer 4:** *Problem description: The problem is to identify (through human help) regions of input space of a classifier such that the classifier confidently makes the wrong prediction.*

    Answer: n/a

5. **Reviewer 4:** *Machine learning: For the most part the paper does not involve machine learning. However, when it describes machine learning (equation 1, definitions 1 and 2) it is at its weakest. This part needs rewriting, see detailed comments below.*

    Answer: See below.

6. **Reviewer 4:** *Results: The results are appropriate and demonstrates the utility of the “Beat the Machine” scheme.*

    Answer: n/a

7. **Reviewer 4:** *However, there is something unsatisfactory from an ’unknown unknown’ perspective. How does one quantify how much of unknown unknowns are the human beings able to reveal to the learner? Are there unknown unknowns that not even human teachers know about?*

Answer: I'm not sure I completely understand this comment. I guess the reviewer is correct that there may be unknown unknowns that BtM can't catch because the humans don't even know. (In the paper the "knower" is the system, not the humans. But certainly the humans may not know some unknown unknowns.) The claim of this paper is that BtM can find unknown unknowns, not that BtM can find all the unknown unknowns. Perhaps philosophically that is impossible. Our experience is that it is striking how much "creativity" people show when challenged—for example, to find unusual forms of hate speech that the firm had not previously even considered. (Sadly.)

8. **Reviewer 4:** *A related question that was not addressed is how the human teachers (the Turkers) learn what are the unknown unknowns with respect to the machine. Did they just randomly guess in the beginning? More importantly, did the humans adapt based on the feedback they received, in order to hone in on the most productive unknown unknown regions?*

   Answer: For these results we intentionally give no guidance to the humans about the internals of the system. Generality, we want humans to probe the system in ways that each one considers the best, in order to tap into the diversity of thinking of each user. We also added some details about the fifth iteration of the design of BTM, that shows that indeed users adapt over time to submit unknown unknowns more accurately. (With the caveat that often these more frequent successes are not as useful as hoped.)

9. **Reviewer 4:** *Domain expert: The paper provides anecdotal evidence from industry users that the Beat the Machine scheme is useful.*

   Answer: n/a

10. **Reviewer 4:** *Lessons: The most valuable lesson in terms of machine learning might be to suggest a possible venue for detecting model mismatch. There is also obviously the practical impact.*

    Answer: n/a

11. **Reviewer 4:** *The paper may give readers the impression that uncertainty-based sampling is a proper active learning strategy – it is not. See Dasgupta & Langford's ICML'09 tutorial* `http://hunch.net/~active_learning/` *I think the argument of unknown unknowns still holds, though.*

    Answer: We rephrased the claim, in a way that is more consistent with the claim about unknown unknowns, and we added a citation to the tutorial.

12. **Reviewer 4:** *Definition 1 is sloppy. It might be cleaner in equation (1) to use $\hat{p_i}$ for the model-estimated posterior $p(y = i|x)$, and then $\hat{ExpCost} = \sum \hat{p_i}\hat{p_j} c_i j$. (It's not clear why $\hat{MinCost}$ needs to be defined in equation (1))). One then defines the true ExpCost using p_i (the true posterior) and $\hat{p_j}$. Furthermore, it is not clear what "high" means – you either define it with rigor, or don't call it a definition.*

    Answer: We have revised the definition of expected and actual costs. Since the true posterior in our case is always a single class, the actual misclassification cost is trivially easy to compute. Regarding the term "high" we would still prefer to keep the current setting: the notion of known-unkowns and unknown-unkonwns is easier to understand when the we treat these regions as a continuum, instead of trying to impose artificial thresholds for defining the regions.

13. **Reviewer 4:** *Same with definition 2.*

    Answer: See above.

14. **Reviewer 4:** *p9 l13: typo*

    Answer: Fixed.

15. **Reviewer 4:** *p10 l29: typo*

    Answer: Fixed.

16. **Reviewer 4:** *p10 l42: typo*

    Answer: Fixed.