CATAPULT Digital

KnowRisk - Ethics Tools Report

Operationalising ethics principles
through the use of applied Ethics tools

June 2021

Table of contents

Section 1
Executive summary

This report focuses on the use of applied AI Ethics tools to operationalise ethics principles within Digital Catapult's technical contribution to the KnowRisk project. The report should be read in conjunction with the Ethics Workstream Report, the scope of which is the ethics work across the KnowRisk project and consortium as a whole.

Digital Catapult's technical contribution to the KnowRisk project falls into two streams of work:

- The further development of an open source Federated Learning library for use by the consortium for privacy preserving distributed machine learning[1]
- The application of the Federated Learning library and a bespoke machine learning model to a specific task relevant to a component of the KnowRisk project: the classification of documents which contain sentences that describe a "risk" or "mitigation"

Alongside this technical work and as part of the Ethics Workstream we selected, adapted, used and evaluated two applied AI ethics tools with the view to enhancing the transparency and robustness of the Federated Learning system. Those tools were:

- **Model Score Cards for Federated Model Reporting**
- **Record On Negative Impact (RONI)**

We believe that this exploratory work has to some extent demonstrated the potential utility of applied AI ethics tools as part of a wider responsible innovation approach. We found the experience of applying these tools alongside our technical contribution to be very useful in terms of expanding our general hands-on experience in applied ethics. This report details the methodology, experimental design, results and final evaluation of these tools in the context of our part of KnowRisk project.

[1] https://github.com/digicatapult/dc-federated

Section 2
Introduction to project context and the use of AI Ethics tools

KnowRisk is a collaborative research and development project. The aim of the project is to develop a platform for organisations to measure, mitigate and price risk for complex modern supply chains. The KnowRisk project utilises AI, distributed ledger technologies (DLT), and geospatial intelligence (GEOINT) to collect, analyse, and verify risk insights. The partner companies involved in the KnowRisk project are SweetBridge, Engine B, Cystellar, Digital Catapult and Intelligent AI, with Sweetbridge being the leading partner.

Given the potential opportunities and risks inherent in such a project, involving advanced digital technologies at different levels of maturity, the application of practical ethics was deemed as essential from an early stage. The Ethics Workstream Report provides a holistic view of this work while this document focuses on the identification, adaption, use and evaluation of applied AI Ethics tools as one aspect of the operationalisation of ethics within the KnowRisk project.

We were able to leverage the past work of the Digital Catapult and the Oxford Internet Institute on a typology of AI Ethics Tools[2] to provide a starting point for tool selection along with a consideration of the initial ethics deep dive results to prioritise needs. The ethics roadmap, detailed in the accompanying Ethics Workstream Report, in particular highlighted areas of specific ethical concern within the consortium: **maintaining robustness of machine learning processes while respecting the privacy of sensitive commercial data and the need for some level of transparency with regards to the machine learning models used and the underlying data.**

To address the above concern and given the distributed nature of both the KnowRisk consortium and the proposed network of users (Insurance companies, small and large businesses connected by supply chains, and technology providers), a Federated Learning approach was adopted. Federated Learning can be defined as “a machine learning technique that trains an algorithm across multiple decentralized edge devices or servers holding local data samples, without exchanging them”[3]. As part of the KnowRisk project Digital Catapult developed an open source library for Federated Learning (FL) that was designed for industrial “cross-silo”, consortium level (<1000 nodes’) deployments[4]. For the purposes of this report it is important to note that we used a standard FL aggregation algorithm called FedAvg which averages the parameters of each of the locally trained models in order to generate a global model at each Federated Learning cycle[5].

Two tools were selected to enhance the privacy, transparency and robustness of Federated Learning systems for use in the KnowRisk platform:

- **Model Score Cards for Federated Model Reporting** (adapted for Federated Learning)

[2] https://arxiv.org/abs/1905.06876
[3] https://en.wikipedia.org/wiki/Federated_learning
[4] https://github.com/digicatapult/dc-federated
[5] https://arxiv.org/pdf/1602.05629.pdf

Model Score Cards for Model Reporting is an established tool for documenting and communicating crucial information about machine learning models to relevant stakeholders in an effort increase transparency and accountability while reducing the risks from information asymmetry and misuse of AI[6]. The model card we created is a living document that describes a machine learning model developed for KnowRisk and was adapted for a Federated Learning context.

- **Record On Negative Impact (RONI)** (adapted for building consortium trust)

Reject On Negative Impact (RONI) was initially proposed as a defence mechanism against various forms of model corruption and data poisoning attacks targeting Federated Learning systems[7]. Our adaptation of RONI, *Record* On Negative Impact, focuses on the context of a small consortium of organisations for which automation of penalties might be unfavourable for consortium cohesion. Therefore the output of RONI in our case takes the form of active Federated model monitoring, recording the impact of model updates from each of the participating parties (insurers) on the global model and it leaves decisions regarding thresholds for negative impact and penalties to the hypothetical parties themselves.

These tools were intended to assist going from actionability, as determined by the ethics roadmap, to operationalisation with use cases from KnowRisk in mind. However due to the early stage nature of the KnowRisk project, this application of two applied AI ethics tools is only intended as a demonstration of how practical tools can be a beneficial aspect of a wider integration of ethics processes and capacity building. Therefore it is worth stating from the outset that there are a number of other tools that could be applied fruitfully to the KnowRisk project as part of a later stage of development and the application of any such tool(s) does not, in and of itself, make a project ethical.

This report will describe how the chosen tools were selected, adapted, used and evaluated. We hope that this example of how tools can form part of a broader engagement with ethics will be a useful case study for others to learn from.

Section 3
Tool identification and methodology

One of the challenges to responsible adoption practices is that whilst the greatest impact (at potentially the lowest cost) can be made right at the beginning of project development, this is also the time with most uncertainty in terms of project definition and scope. The phase of the KnowRisk project covered by this report was indeed an early one - to build a proof of concept - in which the contributions from each collaborator were planned to cohere only towards the end.

Consequently, for pragmatic reasons, the focus of the work on the development and/or evaluation of tools to facilitate responsible technology adoption was on the elements of the KnowRisk solution that were most accessible to the authors - the technical contribution from Digital Catapult (albeit with due consideration of the wider context of the KnowRisk project).

[6] https://arxiv.org/pdf/1810.03993.pdf
[7] https://www.usenix.org/system/files/sec20summer_fang_prepub.pdf

Digital Catapult's technical contribution to the KnowRisk project was to build a proof of concept that showed how Federated Learning can be used to harness private data from multiple parties to build better prediction models while maintaining confidentiality of said data.

The particular prediction model that was demonstrated in this proof of concept analyses text from insurer's risk reports and identifies the risks and mitigations within them. The federated aspect involves training the model separately on private risk reports from (putatively) multiple insurers, and then aggregating the results so as to achieve better predictions than could be achieved from an individual insurer's data alone. This specific prediction task is a component of the overall KnowRisk solution. It should be noted that we identified several other potential applications for Federated Learning within KnowRisk.

The tool selection methodology was to use the ethics roadmap, as detailed in the Ethics Workstream Report, to identify areas of ethical saliency in relation specifically to the Federated Learning proof of concept; to identify tools that might assist in adhering to, or monitoring of, responsible technology practices in those areas; and then to select tools for further investigation.

The following selection criteria were used:

- **Do tools exist?**
- **How well do they address the particular issue identified in the ethics road map?**
- **How mature is the tool, or how readily can it be used?**
- **Can value be added by doing this evaluation?**
- **What functionality is required?**

Further, consideration of the evaluation design (such as if any potential adaptation of the tool might be required, and cost and ease of implementation) was taken into account. The selected tool(s) were then implemented and evaluated for their ability to meet the requirement to mitigate risks or enhance benefits.

Tool Identification and justification

To identify tools that might help to meet the requirements for transparency, privacy and robustness in the Federated Learning setting it was necessary to identify sub-tasks/objectives, before identifying possible solutions to these more specific objectives. For example, the ability to avoid model bias is one element of transparency, and this might be achieved through a combination of careful design, communication of the methodology, agreed standards for data collection, sharing of information about the characteristics of the training data, and ongoing monitoring of model outputs.

In the Federated Learning setting, the responsibility for good practice is complicated by there being multiple participants (as will be the case in any AI supply chain). We therefore mapped the specific tasks / objectives to where in the Federated Learning supply chain they arise or where they need to be addressed (Figure 1, below).

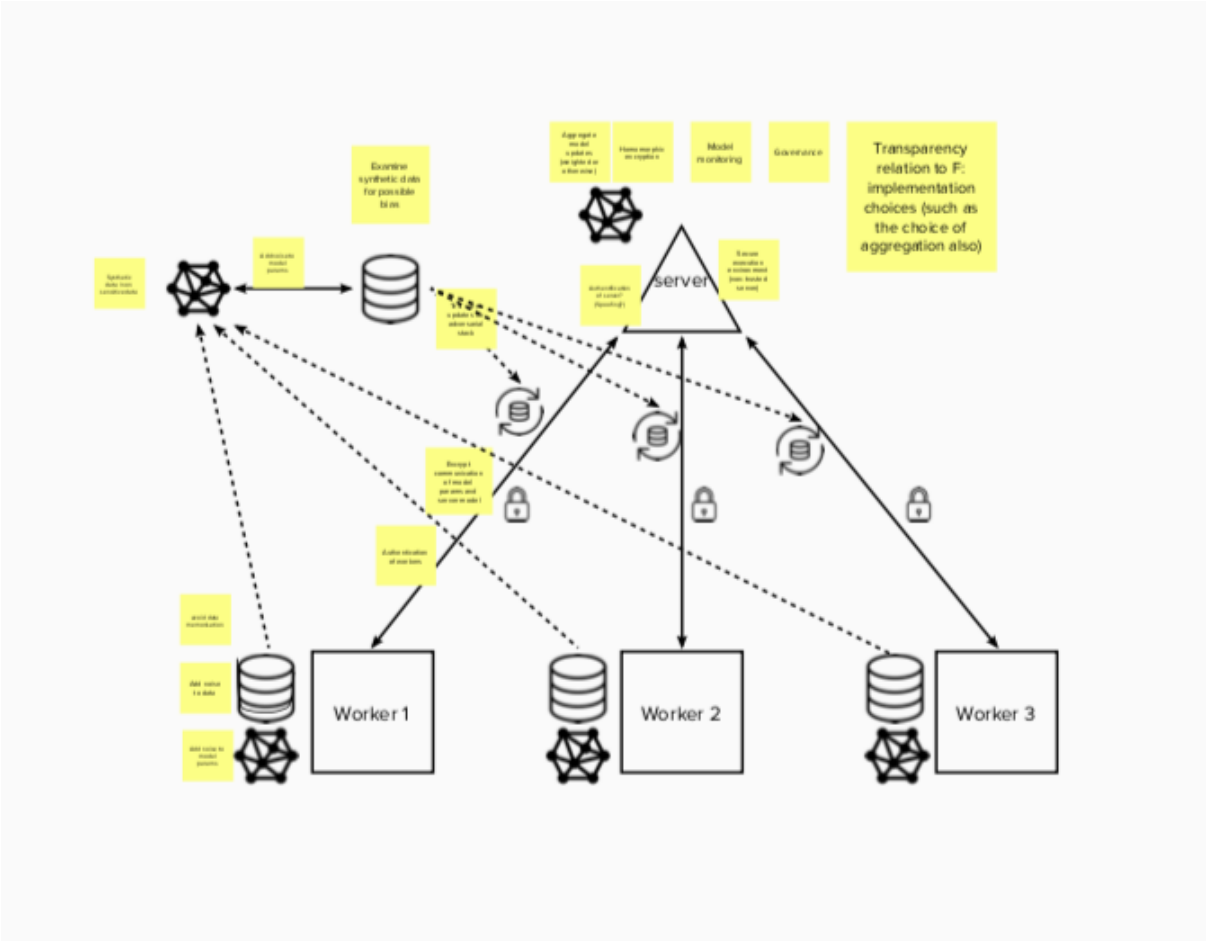


Figure 1: Sketch of Federated Learning system

This allowed research to be targeted at the potential tools that could assist in meeting the specific objectives. We used the AI Ethics Tools Typology[8], a review of relevant literature, and web-search, to identify candidate tools to evaluate against our criteria.

It quickly became clear that the very early stage of the KnowRisk risk prediction proof of concept made tool selection in some areas difficult or nonsensical, since choices were dependent upon any future deployed implementation. Two examples of this are:

- The use of differential privacy to provide privacy guarantees. There has been considerable investment in robust and secure implementations of differential privacy for machine learning (for instance IBM[9], OpenMined[10] etc), to a far higher standard (and trustworthiness) than could be achieved in the resource budget for this work. However, these implementations are typically dependent upon the machine learning model used and/or the machine learning framework deployed, and on the application (in choice of user-selected parameters epsilon and delta). Selection would ideally be made when these elements are fixed, later in the project timescale. In the meantime, it is important to communicate that Federated Learning of itself does not offer any

[8] https://arxiv.org/abs/1905.06876
[9] Diffprivlib: The IBM Differential Privacy Library https://github.com/IBM/differential-privacy-library (As of 16/9, library last updated 26 Jun) (MIT licence - attribution)
[10] PyDP https://github.com/OpenMined/PyDP Openmined's python wrapper for Google's DP C++ DP library 'The library provides a set of ε-differentially private algorithms, which can be used to produce aggregate statistics over numeric data sets containing private or sensitive information"

formal privacy guarantees, but that the use of good regularisation techniques can avoid data leakage through memorisation, and the effort required to reconstruct data from model updates currently makes it rather a theoretical threat, in most situations.

- Fairness; detection of bias. Fairness has attracted a great deal of research interest and a variety of tools and approaches exist to help design fairer systems or identify biases post-hoc. Bias in the risk prediction model (as distinct from a KnowRisk system-level analysis of fairness) would arise from modelling (and scaling) existing biases in human processes or from imbalances or omissions in the training data. Given that the proof of concept was, by definition, a simplified case and one that used synthetically generated training data, it would not be meaningful to evaluate its fairness at this stage. Yet there are distinct challenges to building and monitoring the performance of a Federated System against fairness objectives that will need to be addressed at a later stage. Not least of these is the challenge in understanding the characteristics of the training data used when it is distributed and strictly private. One promising approach to this latter problem is to derive synthetic data sets from the private ones, each with the same statistical properties.

Transparency: selecting Model Score Cards

As a first priority, we chose to focus on transparency since it is fundamental to the investigation of other areas of ethical saliency as well as to the specific collaborative aspects of solution co-development, and on-boarding of customers. In addition, given the early stage of the project, focusing on transparency also allows capturing of proof of concept limitations (and the requirement for further work), such as the privacy and fairness examples discussed above.

Providing accurate information about how the machine learning model was designed, what its purpose is, what data it relies on etc. to participants and stakeholders is the purpose of a number of relatively recent initiatives. These include Partnership on AI's 'AboutML'[11] project ('an ongoing multi stakeholder initiative to enable responsible AI by increasing transparency and accountability with machine learning system documentation), IBM's 'AI Factsheets'[12], Google AI's 'Model Cards'[13] and Microsoft's use of 'Transparency Notes'[14].

Therefore, in terms of the selection criteria we set ourselves:

a. **Do tools exist?** Yes
b. **Closeness of match with particular problem identified in the ethics road map(s)**: These tools seek to increase transparency of models through communicating facts and evaluations relating to their purpose, limitations and design. This should allow informed decision making, and facilitate

[11] Website: https://www.partnershiponai.org/about-ml/
[12] Original Paper: M. Arnold et al, *Increasing Trust in AI Services through Supplier's Declarations of Conformity* (2018); Toolkit (announced July 9 2020): IBM Research AI Factsheets 360
[13] Original Paper: M. Mitchel et al, *Model Cards for Model Reporting*, 2019; Toolkit: Google AI Model Card Toolkit (2020).
[14] For example, https://azure.microsoft.com/en-gb/resources/transparency-note-azure-cognitive-services-face-api/

trustworthiness across the supply chain and with customers and other stakeholders.

c. **Maturity of tool for use**. There is no established standard for documenting machine learning models, but there is much commonality amongst the proposed approaches. Some have been developed into toolkits or part-automated for use in certain circumstances. Integration into existing workflows is lacking, as is methodologies to continuously update the information as models are updated.
d. **Can we add value by doing this PoC test?** We can pilot the use of model reporting in a federated setting and evaluate the ease of use and utility of the tools (against the transparency objective) and publish our results, adding to the know-how and templates available. The primary audience for the specific model information will be other KnowRisk consortium members in the first instance, to assist with understanding and integration into the overall proof of concept. This prototype would inform the design of tooling to achieve transparency in the later, wider, context of deployment, i.e. insurers participating with their private data, and other stakeholders in the KnowRisk system. The outcome will also be of interest to the wider machine learning and AI ethics community, e.g. as a case study for AboutML.
e. **Functionality required?** Identification, and communication, of required information relating to the federated risk prediction machine learning model.

Both Model Cards and AI Factsheets are structured frameworks for reporting facts about machine learning models that have been proposed for widespread adoption. Both are in active development and refinement with various users, but neither has reached de facto standard use. Both have associated toolkits (although Model Cards has a tensorflow dependency) and they are similar in intent and content. Our decision to use Model Cards as the template for reporting was hence made on the basis that this initiative appears to have the most momentum.

Robustness: selecting RONI

The second area of focus relates to the robustness of the risk prediction model, specifically to the observation in the roadmap that: ***"Sometimes, the process of measuring or data collection can distort the thing it wants to measure, sometimes data simply gets distorted at the "entry point". The possibility of "adversaries" that feed data into the system with the explicit intent to distort the outcome can also not always be disregarded."***

In the case of distorted data, it is possible to make some adjustments to the data generation methodology for one or more workers to simulate distortions of data, and hence to test tools that can identify and mitigate against them. Such distortions include biases, data omissions, or processing errors. These are all harder to identify and mitigate in a Federated Learning setting, and are worthy of further research.

There are a number of type of adversarial attacks that can occur within a Federated Learning system for example:

- **Targeted model poisoning**: adversarial worker(s) attempt to manipulate the training process in order to achieve specific aims e.g. targeted misclassification while

maintaining overall model performance (including “stealthy poisoning” to avoid detection)[15].

- **Byzantine failures**: adversarial worker(s) prevent the global model from converging on a reasonable optimum through the introduction of arbitrary model updates (random, drawn from a distribution with higher variance, informed by knowledge of the system). Under normal FedAvg aggregation Federated Learning is not tolerant to even one adversary so mitigations need to be introduced such as dynamically evaluating worker subsets during cost minimization “Krum”[16]. This approach converges in polynomial time and does not need a supplementary test/validation set beyond how the global model is already evaluated. An implementation of the Krum aggregation function is also implemented in IBM’s Federated Learning Library[17].

- **Data poisoning e.g. dirty label data poisoning attacks:** adversarial workers train local models on deliberately corrupted data in a targeted (specific labels are changed) or untargeted manner (labels are to some extent randomly allocated to decrease local model performance and therefore impact the global model)[18].

Overlap with general robustness measures:

- **Class imbalance and bias:** Given that local data is not directly observable this makes efforts to counter class imbalance and potential bias difficult in an FL setting[19]. Tools that attempt to detect adversarial attack by looking at the effect of model updates on the global model may confuse “updates from bad data” with deliberate attacks.

Examples of defences that could be deployed to detect and mitigate some of these attacks include use of more robust aggregation algorithms such as Krum or “trimmed mean” and local update monitoring systems such as Reject on Negative Impact (RONI): Error Rate based Rejection (ERR), Loss Function based Rejection (LFR) and combination of the two used to reject local models[20].

Justification:

a. **Do tools exist?** Yes. There are aggregation function approaches to mitigating against Byzantine attacks and there are further approaches to local model poisoning attacks as well as techniques that can be adapted for federated model monitoring like RONI.
b. **Closeness of match with particular problem identified in the ethics road map(s):** In the context of cross-silo Federated Learning, a tool to monitor for potential attack/failure can increase trustworthiness amongst users and the capacity to act. Given that Federated Learning systems are distributed and opaque with regards to data, a tool to measure and potentially take action based on per-worker/per-party model performance is a good match for the

[15] http://proceedings.mlr.press/v97/bhagoji19a/bhagoji19a.pdf
[16] https://proceedings.neurips.cc/paper/2017/file/f4b9ec30ad9f68f89b29639786cb62ef-Paper.pdf
[17] https://github.com/IBM/federated-learning-lib/tree/main/examples
[18] https://arxiv.org/pdf/1712.05526.pdf
[19] https://arxiv.org/pdf/2008.06217.pdf
[20] https://www.usenix.org/system/files/sec20summer_fang_prepub.pdf

needs identified in the ethics roadmap around consortium governance and auditing of privacy preserving machine learning systems.

c. **Maturity of tool for use.** IBM has quite a complete FL library that includes some implementations of adversarial robustness measures and there are many academic papers available. The use of Digital Catapult's Federated Learning library can make adaptation of tools like RONI easier to implement.
d. **Can we add value by doing this PoC test?** It would be valuable to add robustness features to Digital Catapult's Federated Learning library. Furthermore, there is additional value in demonstrating the mechanisms through which organisations can collaborate with these tools, even if the tools themselves are not novel.
e. **Functionality required?** 1) Monitor the impact of updates from each party (worker node) on the global model 2) Provide actionable insight on potential attacks or failures to users through data visualisation or alert functionality.

There are a number of techniques that provide some level of defence or mitigation against adversarial attack such as the use of more robust aggregation algorithms, however in the context of the KnowRisk project, in which we are not dealing with defending a production Federated Learning system, we decided that a monitoring approach would be a better fit for the needs of the project as explored in the ethics roadmap. This led us to choose Reject On Negative Impact (RONI) as a tool to adapt and implement with a focus on recording anomalies in performance on a per-worker/per-party basis rather than rejecting model updates automatically.

Section 4
Tool adaption and experimental design

Transparency: adapting for Federated Learning

There are a number of additional considerations that we had to factor in when adapting the Model Score Cards for Model Reporting tool for:

1) KnowRisk project overall

KnowRisk is an ambitious and mid-TRL (Technology Readiness Level) applied research project in which the partners are working on interrelated but highly specialised components. The application of Federated Learning to a component of KnowRisk was crucial to a subset of partners (Intelligent AI and Digital Catapult), who were working on a Natural Language Processing (NLP) application for extracting insight from Insurance and risk report documents. For the other partners it would serve as a useful template for applying similar approaches to other components of the project at a later stage. Together with the fact that the intended users (Insurers) of the NLP model were not regularly engaged with the development process, the primary audience for the model score card was the KnowRisk consortium itself - particularly the partners directly involved in developing the machine learning model.

2) The specific machine learning use case

There was fairly limited data available for the proof of concept machine learning model. This necessitated the use of some synthetic data and restricted the scope of the work to producing a demonstrative proof of concept that could be built on at a later stage.

3) Federated Learning

The application of a Federated Learning approach meant that several new pieces of information needed to be provided on the model score card. For example, number of worker node or parties, Federated Learning framework being used, high level information on the data distribution between the parties if that was available. Furthermore, it was decided to document the use of the RONI tool in the model score card so the experimental setup for RONI is included.

The experimental implementation of the adapted model score card was to draft a version alongside the development of the Federated Learning model and share it with the consortium as part of final project outputs to inform future work.

Robustness: adapting for federated model monitoring

Our work took inspiration from a number of sources but primarily a description of the Reject On Negative Impact (RONI) tool for adversarial robustness detailed in Local Model Poisoning Attacks to Byzantine-Robust Federated Learning[21]. In order to implement a version of RONI which *Recorded* rather *Rejected* model update performance we made changes to implementation of the FedAvg aggregation function, used in the DC Federated open source Federated Learning library[22]. The changes enabled the recording of model performance with and without each worker update. We will publish the code for the RONI implementation on the open source repository as part of the KnowRisk dissemination plan.

In the implementation of the NLP risk and mitigation classification model for KnowRisk this meant that for each Federated Learning cycle (when each worker update has been collected and aggregated into a global model) the RONI feature evaluated the performance of a model which only aggregated 3 of the 4 worker updates (a subset model). If one or more of the workers were consistently lowering the quality of the global model compared to the other workers then this could be seen in the relative performance metrics.

The experimental setup for RONI involved creating three datasets which combined ground truth labelled data (risk and mitigation sentences), synthetic data, and “trash” data which is irrelevant data meant to predictably lower the performance of the model that is trained on it.

Datasets:

1) No trash dataset: consisted of a held back test dataset of real risk and mitigation sentences. This dataset was used as a held back test set to evaluate the performance of the subset and global models.

[21] https://www.usenix.org/system/files/sec20summer_fang_prepub.pdf
[22] https://github.com/digicatapult/dc-federated

2) Equally corrupted worker data set. This dataset consisted of two thirds real risk and mitigation sentences and one third “trash” data. This data was split between 4 worker nodes.
3) Unequally corrupted worker data set. This dataset is identical to the equally corrupted dataset except that one of the worker datasets is substituted for a dataset that is completely trash.

It should be noted that in a production Federated Learning setting it may not be possible to have a globally shared test dataset (1 above) due to that inability to access the datasets directly. In practice however, we have anecdotally observed that obtaining limited access to illustrative datasets is common in Federated Learning consortia and is still easier than obtaining full access. Furthermore, it may be possible to generate a synthetic representative test set in a privacy preserving manner using generative networks, this is certainly an area worth exploring in future research.

For full dataset details consult the Model Score Card found in Appendix A.

Modelling details: consult the Model Score Card found in Appendix A.

Experiment:

A Federated Learning system was deployed with four worker nodes and a central server. The machine learning model was trained on local subsets of data before being aggregated by averaging the model parameters (FedAvg).

After each Federated Learning cycle, the RONI feature we implemented would evaluate the performance, (in terms of classification accuracy) with respect to a held out “no trash” dataset, of each subset model (combinations of 3 out of the 4 worker updates) and the global model (an aggregation of all 4 updates).

In order to see if the RONI feature provided useful insights that can improve the robustness of the Federated Learning system, we ran over 10 Federated Learning cycles for both the equally corrupted worker data set (2 above) and the unequally corrupted dataset (3 above).

When the performance curve (accuracy over FL cycles) is plotted for both datasets we should be able to detect whether one of the subset models is consistently better than the rest (because it excludes the corrupted worker).

Section 5
Results

Model Score Card for Federated Model reporting

See appendix A for the full model card.

Record On Negative Impact local (RONI)

Below in Figure 2 and Figure 3 the purple line represents the performance of the global model against the held out test set and the other lines represent the performance of each subset model which exclude a specific worker in the aggregation.

In Figure 2, which shows the results for the equally corrupted dataset, you can see that no one subset model is consistently outperforming the other models with red and green showing similar performance to the global model (in purple). This doesn't indicate that excluding a specific worker increases performance.

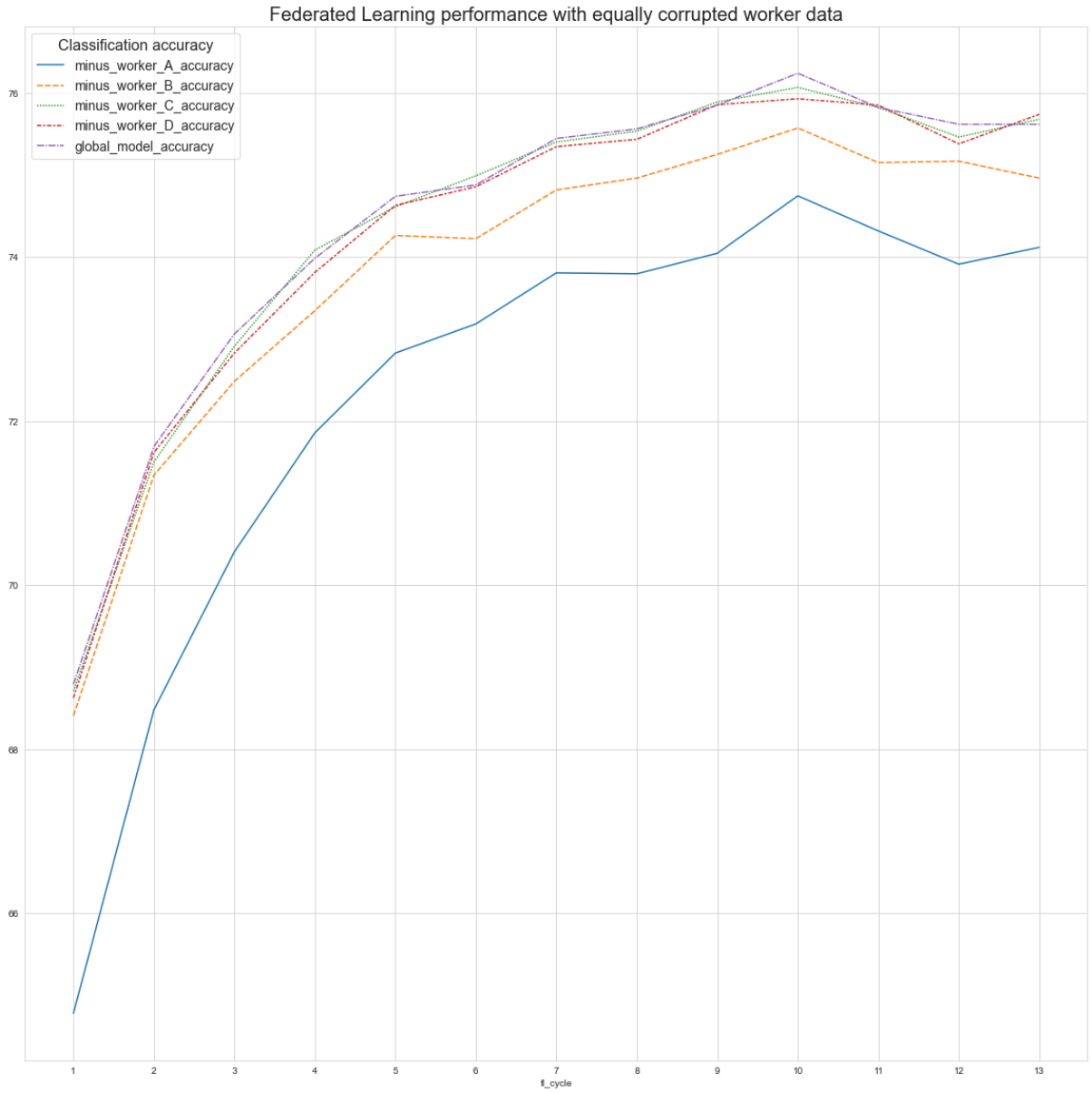


Figure 2: Performance (classification accuracy) on the y axis and Federated Learning Cycle on the X axis for the **equally corrupted dataset**.

Below in Figure 3 you can see that the model subset represented by the red line (excluding worker D) is consistently outperforming the other model subsets and more closely tracks the global model. This indicates that worker D might be worth additional investigation for poor quality data or adversarial threats.

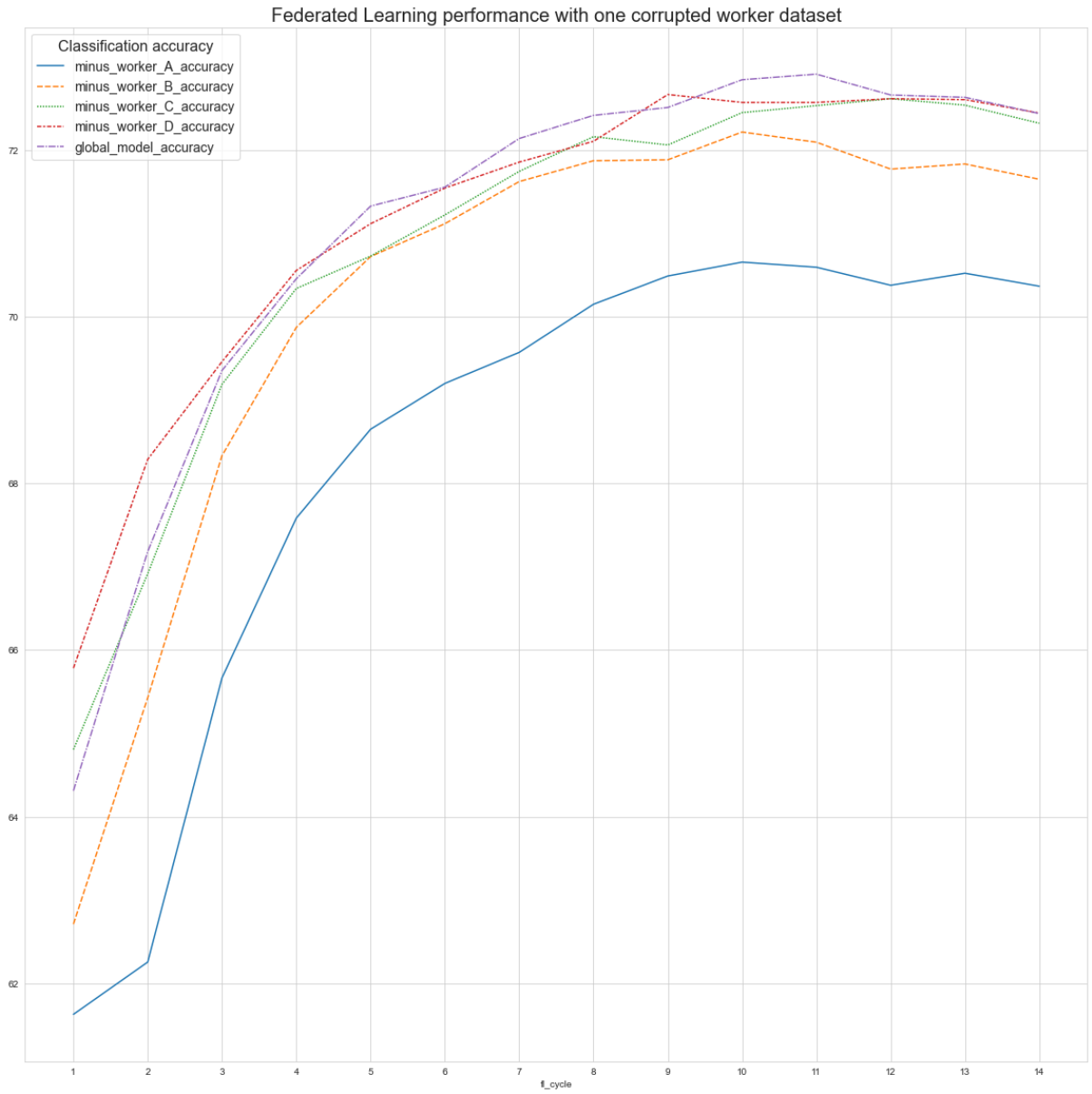


Figure 3: Performance (classification accuracy) on the y axis and Federated Learning Cycle on the X axis for the **unequally corrupted dataset**.

Section 6
Evaluation and discussion

Transparency: Model Score Card for Federated Model Reporting

The first thing to note is that the model card itself does not conform to the template provided in Google's original paper, or to any other template (such as IBM's Fact Sheets). We found the need to tailor the model card to our context - that of a proof of concept model unintended for production, with a relatively limited audience - since we struggled to create a readable and meaningful document by adhering strictly to any template. This is unsurprising, since Google itself does not adhere to its own template in its published model cards[23].

[23] E.g. https://modelcards.withgoogle.com/object-detection

There is clearly a need to tailor model cards to different contexts but that must be balanced by the need to create some consensus about what information a model card must contain, in order to achieve its aims. The information contained in our model card results from the exercise of judgement rather than best practice (since these norms do not yet exist), and are subject to iteration and feedback with its intended audience to best achieve saliency and transparency.

The model has been produced as a 'snapshot' pertaining to the final delivered model for the KnowRisk project. If it has value in communicating between the project participants at this point, that value will disappear as soon as the project evolves and new data, models and integrations occur. In addition, model cards (and their ilk) appear to be most applicable to situations where there is a single model 'owner', and not one in which there may be a) a complex integration of models, potentially each originating from different 'owners' (as is envisaged in the KnowRisk consortium, or other collaborative projects) or b) where federated learning means that the information needed to complete a model card may itself be distributed or may need additional coordination and/or amalgamation.

We can already see that the use of model cards does not fulfil the requirements of transparency and responsibility in the federated setting, even without exposure to potential participants in and/or users of the model (who are currently artificial). This is because the cards are currently static, it is not clear who is responsible for them, nor how to amalgamate distributed contributions. Transparency is especially important because the risk model is anticipated to be only one component of the KnowRisk solution, and in combination, the potential limitations and risks of the solution might become both more opaque and more acute. A holistic communication of solution capabilities and risks will require integration of model cards (or similar) for each component.

Automating reporting tasks where possible will help to integrate them into workflows. Since this is a common problem in machine learning, the use of third party solutions and tools might assist, if they are sufficiently mature and flexible. One promising startup initiative is Parity (getparity.ai) which is a platform, currently in beta, for automating model card workflow including the allocation of tasks and responsibilities, and maintaining flexibility so that users can specify the required information fields.

The resulting model cards must be accessible (and actually accessed!) too. The current model card will be published to a shared private github repository, which at least allows the information to be co-located with the model and associated with a specific version, but it will not be the best choice for everybody. Like Google's model cards, at some point, it may be suitable to make the model card public, via the KnowRisk website, and to think about ways to encourage interaction with it.

Robustness: RONI

What is clear from our experience of implementing and evaluating the Record On Negative Impact (RONI) tool is that, even in a relatively controlled and artificial environment, the results are not concrete enough to inform meaningful alerts let alone automatic mitigations (such as rejecting worker updates).

In a cross-silo consortium setting for Federated Learning, tools like RONI give the parties some means of monitoring model performance that may be indicative of corrupted data without actually needing to see the data itself. This is useful, but such tools can only be part of a wider set of socio-technical solutions - some of which are covered in the wider ethics workstream report.

It is possible that RONI could form part of a suite of Federated Learning tools that include Federated Analytics tools that are able to answer questions about the statistical attributes of distributed datasets without breaching privacy. In order for such tools to be useful, we recommend that they are introduced to participating parties early so that everyone is aware that the capability to detect potential attacks or failures is present. If this discussion is carried out as part of a wider conversation about ethics and accountability, as is occurring as part of the KnowRisk Ethics workstream, then we believe that tools like RONI can contribute to building overall consortium trust in the system. What is crucial is that the application of such tools is not done in a silo and is openly and clearly discussed with all the relevant parties.

Section 7
Conclusion

This report is meant to serve as an open example in which we “show our working out” throughout the process of applying AI ethics tools, from the outset, within an early stage collaborative research project. We are pleased to see that the ecosystem of practical AI Ethics tools has continued to grow, indeed many resources have been added since our tools survey in 2019[24]. As mentioned in the introduction to this report, the application of only two such tools was never meant to be exhaustive but can serve as an illustrative example for the KnowRisk consortium and other practitioners.

With regards to the specific tools we selected, both Model Score Cards for Federated Model Reporting and Record On Negative Impact (RONI) demonstrate some value as tools for increasing transparency and robustness. However, it is also clear that in a production deployment of Federated Learning there is need for a whole system perspective when it comes to implementing infrastructure to support effective model monitoring, privacy, robustness, accountability and also appropriate consortium incentives. This is socio-technical work which should form part of a wider engagement with ethics, which is why this report should be read in the context of the Ethics Workstream report.

We believe that this early and open-ended work in selecting, adapting, implementing and evaluating applied AI ethics tools at such an early stage of a project was a valuable part of the overall Ethics workstream of KnowRisk. Much of the essential ethics work of distilling and communicating ethical values, engaging in critical discussions, as well as interrogating potential risks and benefits has been covered in the Ethics workstream report. What is clear to us is that by drawing a line from the outcomes of that work, for example from the agreed areas of ethical concern in the ethics roadmap, to specific tools that can operationalise ethics in the form of concrete procedures and processes, this ethics tools work can help bridge the gap between the discussion of ethical issues and integrating ethics into the technology itself as well as the human systems that operate it.

[24] https://link.springer.com/article/10.1007/s11948-019-00165-5

Appendix

A: Model Score Card for *Federated* Model reporting v0.2 - RONI experiments

Model Card v0.2

Model Card Date: 01/06/2021
Model Version: 2

Risk prediction

The risk prediction model has been developed by Digital Catapult for the KnowRisk project to classify risk/hazard and mitigation sentences in insurance risk reports.

- **Input:** Natural Language from insurance risk reports. Sentences that were either "risk" or "mitigations" were embedded in dynamically generated documents composed of the sentence of interest and "confounding sentences" from a variety of sources.
- **Features:** Words are first one hot encoded in an n-dimensional vector (where n is the vocabulary size) and then a learned embedding of size m. A document is hence a bag of embeddings. These are averaged to form a single input vector to input to the model for classification.
- **Output:** The model can classify whether a document contains a risk or mitigation. For each document input it returns a single label.
- **Model architecture:** Bespoke implementation of Multi-class Logistic Regression over Bag of Word Embeddings with the embedded dimension hyperparameter set to 10 using the PyTorch machine learning framework. The size of the model depends on the number of unique words in the corpus.
- **Training type:** Federated Learning[25].
    - The model is trained locally for 40 epochs at four (virtual) workers and the model updates are aggregated by a central server before being shared back to the nodes for further local training.
    - A simple ('FedAvg') average aggregation model is used in which model parameters are averaged across all nodes
- **Monitoring:** A '*Record* On Negative Impact' (RONI[26]) feature has been implemented to evaluate the impact of each worker update using a held back validation set. Administrators can set an alert threshold which detects significant irregularities and flags workers for further investigation. The updates are not automatically rejected as in the original Reject On Negative Impact.

Intended Use

[25] McMahan et al, Communication efficient learning of deep networks from decentralized data, 2016
[26] Local Model Poisoning Attacks to Byzantine-Robust Federated Learning

- The model is a proof of concept. It will form one component of a solution that automatically extracts risk and mitigation sentences from insurance risk reports and predicts a risk score for a specific building or set of buildings.
- The intended users are the KnowRisk consortium partners for the purposes of developing and demonstrating a proof of concept application.
- Use by any party other than a KnowRisk consortium partner, for commercial deployment, or use for risk assessment other than specific commercial property insurance related risks is not intended.

Training Data

Training data was generated dynamically by creating documents consisting of multiple sentences, each in the form of a tokenised list of words. One of the sentences was a risk/hazard or mitigation sentence while the remaining sentences were drawn from confounding sources such as a wikipedia dataset. Each document, as a whole, was labelled as a risk or mitigation depending on the label of the risk/mitigation sentence. The sentences contained in the document datasets used in the RONI experiments were from four sources:

Verified labelled data: Ground truth correctly labelled "risk" and "mitigation" sentences shared by insurance partners (321 risk sentences, 389 mitigation sentences).

Augmentation labelled data: Augmenting sentences selected, based on similarity to ground truth data (cosine similarity on BERT embeddings), from publicly available residential risk reports via data.gov.uk[27] (679 risk sentences, 611 mitigation sentences). These sentences were given the label of the ground truth sentence that it was most similar to.

Confounding data to generate input documents: 5000 sentences from the Wiki-Split[28] dataset were used to pad out the document with confounding sentences.

Simulated corruption data ("trash") for RONI experiments: 1000 sentences from the Large Movie Review Dataset v1.0[29]. These sentences were given random labels.

Combinations of these datasets were used to create more or less corrupted datasets distributed over 4 workers for the purposes of testing RONI:

1) No trash dataset: consisted of a held back test dataset of real risk and mitigation sentences. This dataset was used as a held back test set to evaluate the performance of the subset and global models. **(only contains verified labelled data or augmentation labelled data)**
2) Equally corrupted worker data set. This dataset consisted of two thirds real risk and mitigation sentences and one third "trash" data.
3) Unequally corrupted worker data set. This dataset is identical to the equally corrupted dataset except that one of the worker datasets is substituted for a dataset that is completely "trash" **(only containing simulated corruption data)**.

[27]https://data.gov.uk/search?q=fire+risk+assessment&filters%5Bpublisher%5D=&filters%5Btopic%5D=&filters%5Bformat%5D=&sort=best

[28] https://github.com/google-research-datasets/wiki-split

[29] https://ai.stanford.edu/~amaas/data/sentiment/

Evaluation Data

Evaluation data is generated in the same way as the training data. The test-train split is 90% test and 10% train (to offer a harder problem in this artificial experiment).

A distinction should be made between test data that is used in the training of local models, which is split from locally available data and the held back dataset that is used to evaluate subset model performance as part of RONI.

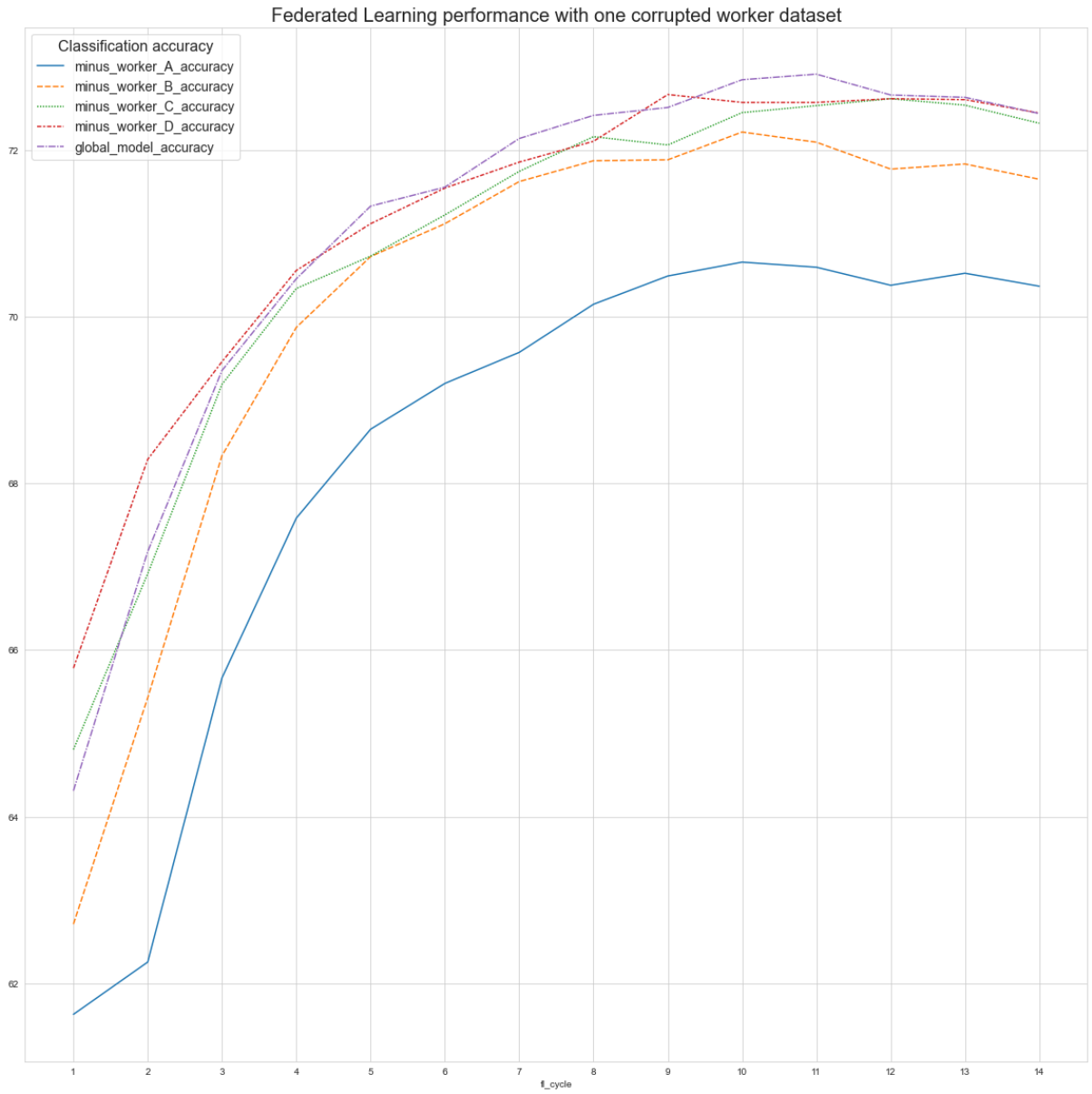
Performance

Model performance is measured using classification accuracy against the evaluation data. Evaluation data is held both locally (normal for Federated Learning) and centrally (not typical for Federated Learning but necessary to test RONI).

Local evaluation is recorded every time a new global model is sent to each node and is tested after 40 epochs of local training.

Evaluation of subsets of the global model (excluding one worker to measure impact) are recorded at each Federated Learning cycle (once all four nodes have sent their local model updates).

Graph of classification accuracy over Federated Learning Cycles:



Limitations and Ethical Considerations

- This model is trained on partially synthetic data which is based on a limited number of risk reports. This model is not intended for use in production as it is not trained on suitably representative data. Real use cases might have much bigger vocabulary and / or classes which lead to bigger data requirements, models, and consequential training and deployment challenges.
- The choice of classes was identified by KnowRisk consortium members to be relevant for a proof of concept, but these may not be a complete or entirely appropriate set for real users and their data. In particular, the problems arising from class imbalance have not been investigated.
- Federated Learning is intended for use in situations where individual contributors wish to keep their data private and secure, but it does not guarantee privacy. There is a risk that data leakage can occur through data memorisation or through reconstruction from model weight updates. In production, use of a range of techniques and technologies are recommended alongside conventional Federated Learning, such as: good regularisation strategies, differential privacy, secure multi-party computation, and homomorphic encryption.

- As with any machine learning model, the training data may contain biases, errors or imbalances that can impact on the efficacy and fairness of the model. In the Federated Learning setting, it might be more difficult to monitor and mitigate against these concerns as the underlying data from each worker is private. For production, further work is required to address these concerns.
- The performance of this model could have a significant impact on the aggregate risk scores that will be generated by the KnowRisk application. Therefore there is potential for harm via biased selection or omission of risk/mitigations which then feed into a biased risk score, the purpose of which is to inform decisions regarding insurance claims. The monitoring and performance metrics for a deployed solution will differ from the simple classification accuracy used here and require further thought.
- This model is intended as both a proof of concept machine learning model to offer specific functionality to the KnowRisk platform (classification of risk and mitigation sentences) and a demonstration of the selected AI Ethics tools in action including this one (Model Score Cards for Federated Model reporting) and Record On Negative Impact (RONI) and so should be viewed as demonstrative and not ready for production use.

Send questions or comments about the model to: ai@digicatapult.org.uk