Co-creating a globally interpretable model with human input

Rahul Nair [1]

Abstract

We consider an aggregated human-AI collaboration aimed at generating a joint interpretable model. The model takes the form of Boolean decision rules, where human input is provided in the form of logical conditions or as partial templates. This focus on the combined construction of a model offers a different perspective on joint decision making. Previous efforts have typically focused on aggregating outcomes rather than decisions logic. We demonstrate the proposed approach through two examples and highlight the usefulness and challenges of the approach.

1. Setting

An interpretable model, one that can be understood by humans, offers several desirable attributes. In high risk application domains (European Commission, 2021) they can address regulatory and compliance needs, offer safe guards against undesirable outcomes, and augment human understanding of an automated decision process.

We consider building an interpretable model for a supervised classification task jointly with humans. This offers a different perspective in aggregated human-AI decision making, in that the focus is on combining decision logic rather than decision outcomes. We leverage an interpretable model that humans can comprehend, and contribute to, through logical conditions.

Ensembling decision logic in this way is motivated by several factors. (a) *Complementarity:* In generating a prediction/outcome, humans leverage broader context and as such observe a different feature space than those present in the training dataset. Further, humans work with an unknown objective that can be complex. To the extent that the decision process can be codified (evidence suggests this is a hard task (Agrawal et al., 2019)), this offers potential to improve machine generated logic since it would be unlikely to learn this from data alone. (b) *Interpretabability:* The joint decision process retains the desirable property of interpretability. (c) *Coverage:* Human inputs that capture changes in regulatory requirements, address situations unseen in training data, or limit undesirable behaviours broaden the scope of the aggregated decision process as a whole.

There are several existing notions of hybrid human-AI decision making and how they are achieved technically. A typical setting is one where the machine handles the 'easy' cases and defers to human experts for the complicated or uncertain cases (Madras et al., 2018). Some view human inputs purely as a post-processing step (Daly et al., 2021), where machine outputs are corrected based on feedback. Several works have formalized aggregation. (Rastogi et al., 2022) provide a unified framework through complementary skills, where the joint decisions are better than those made individually. (Wilder et al., 2020) train a model and delegation decisions jointly. More broadly, there is a growing interest in working logical constraints into ML models (Giunchiglia et al., 2022). (Zhu et al., 2018) offer a broader lens on working in tacit human knowledge into algorithms through value-sensitive design.

In this work, we build on previous work on Boolean decision rules (Dash et al., 2018). They use a mixed integer program to determine a concise discriminative rule between classes. The decision model takes the form of a IF-THEN-ELSE statement, where the IF condition is in disjunctive normal form (DNF), i.e., clauses that are ORs-of-ANDs. The aim of their method is to determine a concise yet accurate rule set that captures training samples. In our extension, human inputs are stated as rules which are included as soft constraints during the search procedure. The optimal rule set therefore combines evidence from the training data and human inputs.

We briefly note the implications of the choice of using logical relations as an interface between machines and humans. In many cases, for example those stemming from regulatory, safety perspectives, or standard business processes, such a representation can be relatively straightforward to derive. In other cases, however, codifying human judgements as rules can be problematic (Agrawal et al., 2019). Moreover, human logic needs to be expressed in terms of the feature space of the training data. This can be a challenge in some cases.

[1]IBM Research Europe, Dublin, Ireland. Correspondence to: Rahul Nair <rahul.nair@ie.ibm.com>.

AI & HCI Workshop at the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

2. Approach

(Dash et al., 2018) formulate the search for an accurate and concise interpretable rule set as a mixed-integer linear programming (MILP) problem.

To describe the formulation, let P denote the set of positive samples, and Z denote negative samples in the training set X. X is assumed to be binarized, i.e., categorical features are one-hot encoded and continuous features are binned. Let K be the set of (exponentially many) possible clauses, namely conjunctions of the binary features in X. Define $K_i, K_i \subseteq K$ as the subset of clauses satisfied by observation i.

The main decision variables are w_k for all k in set K — a binary variable indicating whether conjunction k is selected for the model. Each conjunction k in K has an associated complexity c_k. We take c_k to be the degree of the conjunction, i.e., the number of participating literals. The formulation also defines ξ_i for $i \in P$ (i.e., for all positive samples) to indicate incorrect classification, i.e., a false negative.

The objective seeks to minimize total Hamming loss, where the Hamming loss for each sample is the number of conjunctions that must be added or removed to classify it correctly. Specifically, this is expressed as

$$\min_{\xi,w} \sum_{i\in P} \xi_i + \sum_{i\in Z} \sum_{k\in K_i} w_k. \tag{1}$$

The first term represents the false negatives and the second term the false positives for a choice of conjunctions. False positives add more than 'one unit' if they satisfy multiple selected conjunctions, all of which must be removed to classify the instance correctly. This objective is subject to constraints:

$$\xi_i + \sum_{k\in K_i} w_k \geq 1 \quad \xi_i \geq 0, \qquad \forall i \in P \tag{2}$$

$$\sum_{k\in K} c_k w_k \leq C \tag{3}$$

$$w_k \in \{0,1\} \qquad \forall k \in K. \tag{4}$$

Constraint (2) states that for each positive sample, we either have a false negative ($\xi_i = 1$) or include a rule that correctly represents this observation (i.e. a conjunction from the set K_i). Constraint (3) bounds the total complexity of the selected rule set by a parameter C. Constraint (4) restricts the decision variables w_k to be binary.

Problem (1)–(4) is intractable as written, even with advanced MILP solvers, because the set K is very large and it is prohibitive to generate the entire set of conjunctions. In any case, only a few w_k tend to be selected in the final solution. The authors (Dash et al., 2018) use a column generation (CG) procedure, which is an iterative algorithm by which candidate conjunctions are generated at each iteration only if they can improve the overall objective. The method to generate a new candidate is called the pricing problem and the original model (1)–(4) is called the master problem. The CG procedure can be summarized by the following steps:

1. Restrict the master problem to a small subset of conjunctions $J \subset K$ and solve its linear programming (LP) relaxation obtained by relaxing the constraint $w_k \in \{0,1\}$ to $w_k \geq 0$.
2. Solve the pricing problem to find conjunctions omitted from J that can improve the objective. Add these conjunctions to J.
3. Repeat steps 1 and 2 until no improving conjunctions can be found.
4. Solve the unrelaxed master problem ($w_k \in \{0,1\}$) restricted to the final subset J.

We refer to (Dash et al., 2018) for the formulation of the pricing problem and more details in general.

2.1. Human input

Assume some rules for the task are known, either through domain-knowledge, business process rules, regulatory criteria, or safety considerations. As an example, in a mortgage approval task, a condition

$$(\text{LTV} \geq 90\%) \vee (\text{LoanAmount} \geq 3.5 \times \text{Income})$$

can be used to reject an application based on Loan-to-Value (LTV) ratio and limits on loan amounts set by regulators.

Take U, $U \subset K$, to be a set of known conjunctions. We modify objective (1) as

$$\min_{\xi,w} \underbrace{\sum_{i\in P} \xi_i + \sum_{i\in Z} \sum_{k\in K_i} w_k}_{\substack{\text{Machine objective} \\ \text{(Hamming loss)}}} + \underbrace{c_u n \sum_{k\in U} (1 - w_k)}_{\substack{\text{Human inputs} \\ \text{(violation penalty)}}}. \tag{5}$$

This aggregated objective can be interpreted as follows. The model imposes a penalty each time a user-provided input rule is not selected. The latter term serves as regularization and c_u is the lagrange multiplier as a fraction of the dataset size n. Since the first two terms in the objective represent the Hamming loss, $c_u n$ can be interpreted as the additional Hamming loss that is incurred before a user provided constraint is dropped from the model.

A variant that merits consideration is when human inputs are not known precisely. This occurs when the knowledge of the task is incomplete or rules are only partially known.

Denote U' as a set of such partial conditions. This set can also be viewed as a template. We define a distance metric $d(k, U')$ which computes how similar a conjunction is to provided templates. The objective (1) can be rewritten as

$$\min_{\xi,w} \sum_{i\in P} \xi_i + \sum_{i\in Z} \sum_{k\in K_i} w_k + \underbrace{+c_p \sum_{k\in K} d(k, U') w_k}_{\text{distance penalty for partials}}. \quad (6)$$

This penalizes conjunctions that are not like those provided in the template set. We note, the parameter c_p does not have the same interpretation as c_n in (5). The distance metric $d(k, U')$ essentially compares a conjunction to a set of conjunctions. The metric can be computed either by comparing rule semantics or by statistical means if a supporting dataset is available (Nair et al., 2021).

2.2. Evaluation

The resulting model should generalise well and offer suitable interpretability for domain users. Evaluating a model comes with some subtle challenges in this setting. For generalisation, typical metrics like test accuracy work. However, one can only draw conclusions if there is supporting data. If the rationale for human inputs is to capture conditions not available in the data, then test data sets may also not reflect such conditions. This occurs frequently in practice, e.g. changing business conditions (increase in LTV values to qualify for a mortgage). Standard generalisation metrics will not offer the full picture in these cases.

Interpretability is generally considered as being the length of the rule set. Shorter rules are preferable and considered more interpretable. One additional measure of interpretability in this setting is related to the content of the rules. If domain-specific inputs are known then one measure is how well the rules mimic domain-specific semantics that are familiar to experts. Such a measure would penalise rule sets using negations for example. We use a semantic rule similarity proposed in (Nair et al., 2021) that solves an assignment problem to determine the least cost mapping between two rule sets. A rule similarity of 1.0 indicates perfect agreement in semantics and a value 0 indicates no common semantics.

3. Examples

3.1. Tic-Tac-Toe

The key ideas are first shown on the game of tic-tac-toe. Given the state of the end board[1], consider the task of predicting if 'x' wins the game. In tic-tac-toe, there are exactly eight rules under which a model predicts `true`: three 'x' verticals, three horizontals, and two diagonals (Figure 1). Irrespective of the state of the rest of the board, if any of these eight conditions occur, the prediction must be true. This is a noise-free classification task with deterministic rules.

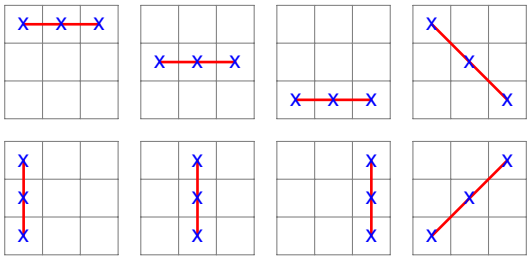

Figure 1. The 8 deterministic rules that classify 'x' as winning a tic-tac-toe game.

We aim to learn these rules from data on the end board configuration under varying levels of human inputs. While the rules are picked randomly, they are assumed to be known perfectly. As (Dash et al., 2018) report, their algorithm is able to recover all rules from this dataset. We therefore limit the data available to the algorithm as well, to degrade machine performance for illustrative purposes.

We implement the model described by objective (5) and constraints (2)–(4) and use CPLEX as the solver. We measure the test accuracy (Figure 2) and semantic rule similarity (Figure 3) on a hold out set over five folds. The complexity parameter C is set to 24, and $c_n = 0.05$. We vary the amount of training data available to the method along with extent of human input, i.e. either no inputs, all 8 rules or partial rules of 2, 4 or 6 randomly sampled rules.

The effects are most pronounced when the algorithm has access to only 5% of the data. Here, an machine-only solution has a median accuracy of 59%. The algorithm generates a perfect model for the case when the human inputs all 8 rules with perfect information about the game. When the rules are partial, results fall in between these two extremes.

Rule semantics present a similar story shown in Figure 3. Machine-only generated rule sets exhibit poor semantic similarity to the 8 known win rules and improve with additional human-provided rule sets. However, even in this simple case, without significant human-guidance, learnt rules are expressed differently even if they have the same implication. For instance, the diagonal rules are often expressed as the absence of an 'o' rather than a presence of an 'x'. This illustrates that the human-AI model outperforms the machine-only outcomes in generalisation and rule semantics.

[1] https://archive.ics.uci.edu/ml/datasets/Tic-Tac-Toe+Endgame

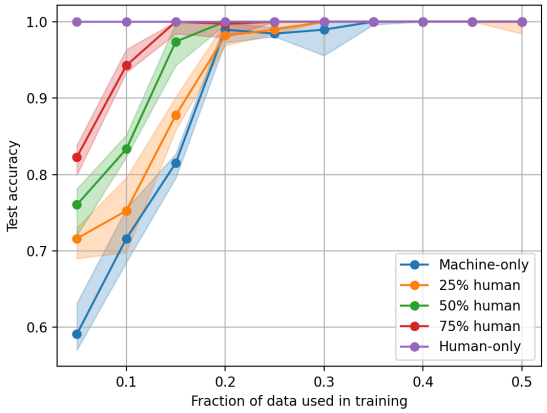

Figure 2. Median test accuracy (with 25th and 75th percentiles) from a 5-fold cross-validation, shown as a function of training size and size of human inputs for tic-tac-toe end game.

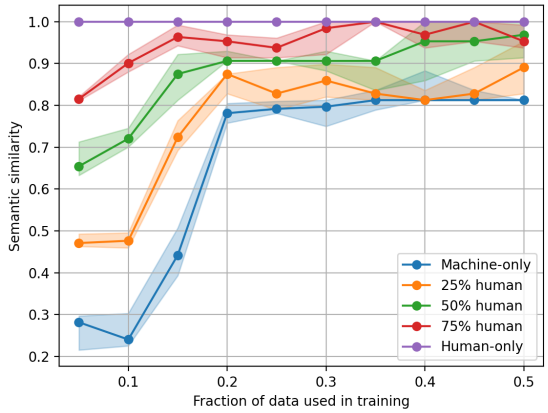

Figure 3. Semantic rule similarity (with 25th and 75th percentiles) from a 5-fold CV. Higher is better.

3.2. Healthcare

We now consider a more practical case in healthcare. Sepsis is a serious medical condition due to organ dysfunction brought about by infection (Singer et al., 2016). It is associated with fast progression of the condition and a high risk of death. Rapid treatments are a key way to better manage the condition and reduce mortality risk. Several aspects of sepsis are not very well understood (Hotchkiss & Karl, 2003).

In practice, scoring systems like SIRS and qSOFA are used to assess if sepsis likely so treatment can be initiated quickly. These methods use logical conjunctions based a patient's vital signs. qSOFA suspects sepsis if two of the three conditions on respiratory rate ($\geq$ 22 breaths per minute), blood pressure ($\leq$ 100 mmHg), or altered mental state are true.

Treatments vary by severity of sepsis which is generally classed in three categories (sepsis, severe sepsis, and septic shock). Several drug regimes exist to treat sepsis. While anti-bacterials target source of the infection, mild and severe sepsis are each treated with different medications. Drugs meant for severe sepsis cannot be administered before those used for mild sepsis, as this can lead to antibiotic resistance (Li et al., 2020). Another class of drugs (vasoactive and diuretics) are used to reduce severe symptoms. These considerations form the basis of expert human input.

We consider the task of predicting survival condition of patients using 30-day mortality outcomes. From the MIMIC-III (v1.4) dataset (Johnson et al., 2016), we extract data on ICU admissions with suspected sepsis (1,783 admissions). For these admissions we retrieve age and gender of the patients along with several vitals such as respiration, temperature, etc. We include medications relevant to sepsis that were administered over the course of their ICU stay. Similar to (Li et al., 2020), we consider 3 drugs for severe sepsis, 7 for mild sepsis, 3 vasoactive, and 1 diuretic medication.

Our setting differs from that of (Li et al., 2020) in that the temporal dynamics are simplified in the following manner. Abnormalities in vital signs are aggregated over a day. Missing vital signs are considered to be normal. The survival condition is measured at 30 days after hospital admission.

Human input For our experiments, we consider the conditions on administering drugs in a specific order similar to (Li et al., 2020), i.e.

$$\text{Survival Outcome} \leftarrow \text{Use Drug}_1 \wedge \text{Use Drug}_2$$

where, Drug_1 denotes set of drugs used for severe sepsis, and Drug_2 are drugs for mild sepsis.

Three models were trained. One 'machine-only' model, a second rule induction with human input, and a third non-interpretable model as a baseline. The first two were run for several values of the complexity budget $C = \{5, 10, 15, 25, 30\}$. The non-interpretable model is a Random Forest from scikit-learn with all default values (Pedregosa et al., 2011). All models were run over 5 cross-validation folds. The two metrics of interest, test accuracy (Figure 4) and rule similarity (Figure 5) computed. Rule similarity here is relative to human-input.

The performance of both interpretable models is better than the baseline RF model for most cases. The human-assisted variant does no worse than the machine-only variant. When $C = 25$ it offers some gains. Despite the performance being similar, rule similarity relative to human-input is better.

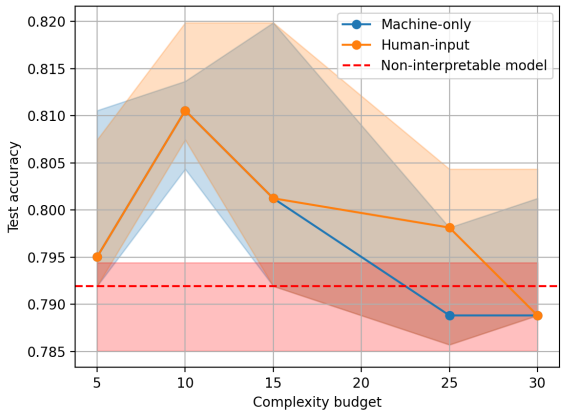

Figure 4. Median test accuracy (with 25th and 75th percentiles) from a 5-fold cross-validation

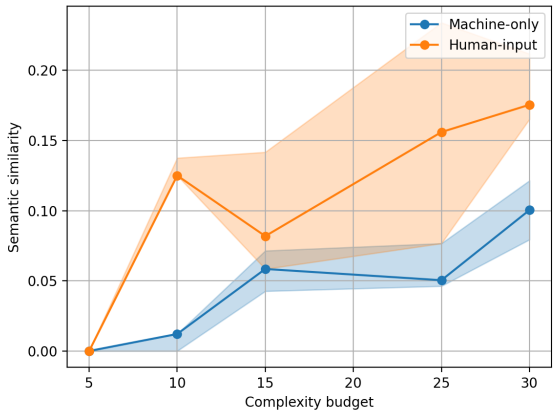

Figure 5. Semantic rule similarity (with 25th and 75th percentiles), 5-fold CV showing

What does this mean? On one hand it is somewhat expected, as human expert rules were provided as input to one and not the other. However, considering similar performance, this suggests that the human-input variant used domain-specific logic instead of statistical basis to construct its rule set. The main advantage therefore is that the resulting model reflects domain-expert intuition and insight.

A sample rule set for the best performing model when $C = 10$ which achieves 82.9% accuracy is shown below.

$$\begin{aligned}\text{Positive Outcome} \leftarrow\ & \\ & (\text{Use Vancomycin} \wedge \text{Use Metronidazole}) \vee \\ & (\neg(\text{severe+mild}) \wedge \neg(\text{Use Furosemide}) \wedge \\ & \text{Normal CRP} \wedge \text{Abnormal Creatinine} \wedge \\ & \text{Use Norepinephrine} \wedge \text{Abnormal Lactate} \wedge \\ & \neg(\text{Use Metoprolol}) \wedge (\text{Age} > 52.0))\end{aligned}$$

The rule set consists of two conjunctions. The first is from human-inputs and refers to the drug combination for patients with severe and mild sepsis. The second refers to a cohort with the same severity throughout their admission and specific vitals, e.g. abnormal lactate, and with specific medications. We caution that this rule set is illustrative only and without clinical validation.

4. Discussion

A model to generate a joint interpretable model with human input has been proposed and demonstrated on two examples. The main idea explored is the combination of decision logic directly, as opposed to aggregating outcomes. Aggregating decision logic in this way offers pipelines that reflect expert insights. By construction, the rule sets have bounded complexity and is fully interpretable. This makes it useful for practical applications.

Disagreement: When human inputs directly contradict the data, the optimization model may prefer one or the other depending on the parameter c_u. Larger values will make human provided rules more 'sticky' in the final model. Preliminary results suggest this is the case; however, it is worth considering other ways to express such cases within the model.

Complementarity: In some cases, it may be feasible to assess how complementary human and machine inputs are, as done in (Rastogi et al., 2022). The main constraint is the availability of evidence to make such an assessment. If it is available, then the manner in which the complexity budget C is shared between human and machine generated conjunctions can provide additional insights.

Satisfiability: We have explored the case where human inputs are treated as soft constraints, i.e., can be violated. In cases where such inputs need to be satisfied, these can be treated as hard constraints, by including constraints in the optimization model directly. This strictly enforces human inputs.

Additionally, we have touched upon some challenges associated with such an approach. Chief among them is the codifying of human judgements in terms of available attributes. Empirical challenges around evaluations can also be challenging if human inputs are based on exogenous factors not reflected in training data. The use of logical conjunctions in the DNF form uses axis-aligned primitives. This is a natural and flexible approach but has limits when more advanced feature transformations need to be expressed.

5. Acknowledgements

Thanks to Jonathan Epperlein and Anne-Marie Cromack for comments and review of an earlier draft of the paper. This work was partially funded by the European Union's Horizon Europe research and innovation programme under grant agreement no. 101070568.

References

Agrawal, A., Gans, J. S., and Goldfarb, A. Exploring the impact of artificial intelligence: Prediction versus judgment. *Information Economics and Policy*, 47:1–6, 2019.

Daly, E. M., Mattetti, M., Alkan, O., and Nair, R. User driven model adjustment via boolean rule explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7):5896–5904, May 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/16737.

Dash, S., Günlük, O., and Wei, D. Boolean decision rules via column generation. In *Advances in Neural Information Processing Systems*, pp. 4655–4665, 2018.

European Commission. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. 2021. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

Giunchiglia, E., Stoian, M. C., and Lukasiewicz, T. Deep learning with logical constraints, 2022. URL https://arxiv.org/abs/2205.00523.

Hotchkiss, R. S. and Karl, I. E. The pathophysiology and treatment of sepsis. *New England journal of medicine*, 348(2):138–150, 2003.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Li, S., Wang, L., Zhang, R., Chang, X., Liu, X., Xie, Y., Qi, Y., and Song, L. Temporal logic point processes. In *International Conference on Machine Learning*, pp. 5990–6000. PMLR, 2020.

Madras, D., Pitassi, T., and Zemel, R. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in Neural Information Processing Systems*, 31, 2018.

Nair, R., Mattetti, M., Daly, E., Wei, D., Alkan, O., and Zhang, Y. What changed? interpretable model comparison. In Zhou, Z.-H. (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 2855–2861. IJCAI, 8 2021.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Rastogi, C., Leqi, L., Holstein, K., and Heidari, H. A unifying framework for combining complementary strengths of humans and ml toward better predictive decision-making. *arXiv preprint arXiv:2204.10806*, 2022.

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., Chiche, J.-D., Coopersmith, C. M., et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8):801–810, 2016.

Wilder, B., Horvitz, E., and Kamar, E. Learning to complement humans. *arXiv preprint arXiv:2005.00582*, 2020.

Zhu, H., Yu, B., Halfaker, A., and Terveen, L. Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–23, 2018.