Lessons Learned from Problem Gambling Classification: Indirect Discrimination and Algorithmic Fairness*

Christian Percy[1], **Artur d'Avila Garcez** [2], **Simo Dragicevic** [1], **Sanjoy Sarkar** [1]
[1] Playtech Plc
[2] City, University of London
chris@cspres.co.uk, a.garcez@city.ac.uk, Simo.Dragicevic@playtech.com, Sanjoy.Sarkar@playtech.com

Abstract

Problem gambling is a public health issue with approximately 300,000 individuals suffering harm in England and 1.5 million at risk. Many gambling operators rely on Machine Learning (ML) algorithms to identify online players at risk. Models are typically gender-blind (gender not included as an input), reflecting the sensitivity of protected characteristic data. However, some stakeholders worry that gender continues to influence the model via other variables (indirect identification) and worry about differential model performance by gender (algorithmic fairness). In this paper, we investigate these concerns using real-world data from 22,500 players across two gambling operators. We propose a method for testing the indirect identification of a protected variable. We identify near-zero levels of indirect identification of gender. Regarding algorithmic fairness, a slight pro-female bias is found in the first ML model and a moderate pro-female bias in found in the second ML model. The challenge is to mitigate such bias without the intrusion of compulsory gender data collection. We propose a new approach which uses gender data for training only, constructing separate models for each gender and combining trained models into an ensemble that does not require gender data once deployed. Since harm identification adopts a precautionary principle, if any one model indicates potential harm, the player is flagged as at risk. This approach is shown to reduce the difference per gender in the True Positive Rate (TPR) of the models from 7.2% points to 4.0% points. This is shown to be better than what can be achieved by simply altering the models' classification thresholds. Both the indirect identification and the algorithmic fairness approaches are part of a wider framework and taxonomy being proposed towards the ethical use of Artificial Intelligence (AI) in the gambling industry.

Introduction

Problem gambling is a public health issue with approximately 300,000 individuals suffering harm in England and 1.5 million at risk. Many gambling operators rely on Machine Learning (ML) algorithms to identify online players at risk. Models are typically gender-blind (gender not included as an input), reflecting the sensitivity of protected variables. However, protected variables may continue to influence an ML model outcome by proxy (via other variables) in ways that can make the identification of bias even harder, and make the bias correction towards algorithmic fairness impossible.

In this paper, we report the lessons learned from work by Playtech plc, a provider of B2B and B2C gambling software services, investigating the role of gender in Playtech's gambling harm identification algorithms. The paper connects cross-sector concerns around algorithmic fairness with the specific public health work on problem gambling mitigation. In online gambling, there is a need to balance harm protection against e-consumers' desires and rights to privacy, with a default that minimises the use of sensitive data. As such, online gambling is a relevant and challenging test case for exploring issues around algorithmic bias. The main contributions of this paper are:

- A technique for identifying when indirect discrimination exists in ML models where a potentially problematic variable has been dropped ('model-matched indirect identification').
- An approach for incorporating insight based on protected variables into the classification algorithms without requiring such data to be collected compulsorily from individuals ('blind-separate models').
- The evaluation of the above technique and approach using gender in a real-world use case in the gambling sector, which reveals near-zero indirect identification and yet the potential to improve algorithmic fairness, as defined by a reduction in the difference between the model's true positive rate for men and women.

We see this work as part of a broader project, both within the gambling sector and on the ethics of AI more generally. In the gambling sector, this paper offers the basis for the formation of a sector-wide working group to study algorithmic fairness and to consider how to address multiple objectives, such as overall true positive rates versus true negative rates, model performance disparities by key customer groups, model aggregation policies, parsimony and transferability.

On the ethics of AI more generally, a framework and taxonomy is being developed, which includes the concerns of algorithmic fairness and bias identification and other use

[*] The authors would like to thank Charmaine Hogan and Lauren Iannarone at Playtech for their constructive challenge and feedback.

cases to improve fairness. Our approach recognises that the diversity and complexity of ML models and data sets, use cases and stakeholders' priorities are such that no single technique can be recommended universally, but also certain principles and a framework should be sought to be developed and applied specifically and across the gambling industry given the importance of the issues being discussed[1].

The remainder of the paper is organised as follows: in Section 2, we position the paper in the context of the related work; in Section 3, we describe the proposed technique and fairness approach within the problem gambling use case; in Section 4, we evaluate the influence of gender on indirect identification and options to enhance algorithmic fairness; in Section 5 we conclude and discuss directions for future work.

Related Work

Algorithms for problem gambling harm reduction

Problem gambling is a public health issue with 300,000 estimated individuals in England self-reporting as experiencing harm (0.7% prevalence, Gambling Commission, 2018) and a further 1.5 million thought to be at risk. Sector organisations take a range of steps to mitigate, identify and intervene to reduce gambling-related harm, including the development of Machine Learning algorithms to identify players at risk of harm. Some examples of early interventions that operators might take, having identified someone above a certain threshold of risk or possible harm, include tailored responsible gambling messages or reduced marketing activity.

Playtech plc has an in-house suite of ML algorithms trained to identify players with similar characteristics as players who have self-identified as experiencing harm - see Percy et al (2016) for background on these supervised ML models. Several of Playtech's operations fall under the purview of the European Union's General Data Protection Regulations (GDPR; officially adopted in April 2016). Adopting a precautionary interpretation of GDPR principles of data minimisation and protections for special category data, the default decision on algorithms implemented by Playtech was not to incorporate the player's gender, achieving typical cross-validation AUROC rates of 95%+ on balanced data sets by using behaviour and transaction data alone.

However, the adoption of gender-blind algorithms is under review. Regulatory advice from the UK Gambling Commission in 2018 suggests that demographic data can be used as part of satisfying regulatory requirements[2]. There is also increasing awareness of the role of gender in problem gambling. A UK gambling charity has reported that the rate of problem gambling amongst women increased by a third in the preceding five years to 2019, a faster rate of increase compared to men in the same period (15%)[3]. Therefore, as the industry becomes increasingly reliant on ML algorithms to detect problematic play, the changes in demographic profiles of problem gamblers raises questions on the suitability of historic gender-blind data sets, their role in training models, and their potential impact on model efficacy. This can be seen as part of a broader trend arguing that the traditional gender-blind approach in gambling research is inappropriate (Baggio et al, 2018) and implicitly male-biased to the detriment of female gamblers (McCarthy et al, 2019; Venne et al, 2019).

The research reported in this paper was initiated to address two possible stakeholder concerns that point towards opposite modelling responses. The first is that gender remains an (unwanted) influence on the gender-blind model via its indirect associations with other variables (indirect discrimination). Here, the goal is to remove as much of this influence as possible. The second concern is whether there is a missed opportunity for using gender data in a way that enhances algorithmic performance and fairness by identifying and mitigating differences in model performance by gender group (algorithmic fairness). The first concern is motivated directly by an awareness of the sensitivity of gender data, both to consumers and in legislation (GDPR). The second concern reflects an awareness of average structural differences between men and women that may be relevant for predicting gambling risk. For instance, research has related testosterone levels to risk-taking and pathological gambling (Stenstrom and Saad, 2011), identified gendered behavioural patterns in gambling problems (Wong et al, 2013), and observed gendered patterns in the types of online behaviour that can be addictive (Su et al, 2020).

Bias in AI algorithms and mitigation

Concerns about bias in AI algorithms in relation to sociodemographic traits have now become widespread. The second of Google's seven principles for AI is to avoid creating or reinforcing unfair bias[4]. Organisations and researchers are responding to this concern in different ways, which can be grouped based on whether they seek to intervene at the input level, at the model level or at the output level.

At the input level, one approach adopted, already discussed, is to exclude the variable corresponding to the sociodemographic trait in question. For instance, Goldman Sachs in its operation of Apple Card deliberately avoid collecting and using data on sensitive characteristics such as gender, race or age, using this approach to defend against concerns

[1] We seek to respond to calls by AI researchers, such as Goodman (2016) of the Oxford Internet Institute, to develop frameworks for so-called 'algorithm audits', and sector bodies, such as the EC's Advisory Committee on Equal Opportunities for Women and Men (2020) which recommends monitoring of algorithms for discrimination and further work, calling for work to develop and share good practices. The UK Government's CDEI further describes a lack of clear regulatory standards and quality assurance (e.g. around algorithmic bias) as one of the five key trust-related barriers holding back AI (CDEI, 2020:4).

[2] https://www.gamblingcommission.gov.uk/for-gambling-businesses/Compliance/General-compliance/General-Data-Protection-Regulation-GDPR.aspx

[3] www.telegraph.co.uk/news/2020/01/15/female-gambling-addicts-growing-faster-men-amid-rise-online (accessed August 2020)

[4] https://ai.google/principles/ (accessed August 2020) (published 2018 at https://www.blog.google/technology/ai/ai-principles/)

of gender bias[5]. However, this practice typically proves an insufficient defence in the face of evidence of gender bias in the outcomes - New York's Department of Financial Services opened an investigation into Apple Card in late 2019 given different credit limits provided to men and women despite apparently similar financial circumstances[6]. In another example, analysis by Obermeyer et al (2019) revealed that the use, in a widely-used commercial algorithm, of health costs as a proxy for healthcare needs resulted in anti-Black racial bias; the authors recommend removing health costs as an input variable as a proxy for needs. Another approach is to increase the availability and diversity of training data relating to the input variable in question, which was part of Microsoft's 2018 strategy for reducing the error rate discrepancy between men and women and between lighter skin tones and darker skin tones in its image classification tool Face API[7].

At the model level, algorithms or their parameters can be adjusted to reduce the extent to which a model draws on certain patterns in the input data. One example of this is the gender-debiasing techniques developed for word embedding solutions (Bolukbasi et al, 2016), noting that the authors describe a mixture of adjusting inputs and model-level adjustments.

At the output level, Moerel (2018) describes LinkedIn's recruitment tool as a way of enforcing quotas using the rankings produced by an algorithm in order to match a predefined desirable ratio. The tool can subdivide candidates by gender, rank each candidate within each gender using its algorithm and then put forward an equal number of men and women to the hiring manager for consideration.

Some of the techniques above have come under challenge. For instance, the simplistic approach of dropping socio-demographic input variables (blinding an algorithm) has come under challenge for inadvertently distracting from fairness by reducing visibility of the issue, by ignoring possible proxy variables for socio-demographic traits and by ignoring opportunities to implement other solutions - see, e.g. the analysis of US College admissions by Kleinberg et al (2018) which argues for data-led proactive intervention at the output level.

Focusing particularly on identifying output-level bias, new tools are being developed to identify whether ML algorithms are biased in terms of having systematically worse performance (e.g. lower accuracy) for particular groups. Facebook announced the testing of an internal tool to do this in 2018, Fairness Flow, which was discussed further in its July 2020 Civil Rights Report as part of efforts to tackle algorithmic discrimination[8]. Google's open-source What If Tool in TensorBoard launched in 2018 to help ML developers to visualise differences in classification from key variables, identify borderline cases for particular classifications and explore the impact of counterfactuals as part of assessing whether an inappropriate social bias might have been absorbed from the training data or otherwise reflected in the model[9]. The use of counterfactuals for explainable AI, e.g. White and Garcez (2020), has become increasingly associated with the goals of fairness in ML. Various other methods addressing fairness which have been proposed recently, have adopted their own measures of fairness. Notably, Dwork et al (2012) introduces a framework for fair classification by a task-specific metric for maximizing utility subject to a fairness constraint. Agarwal et al (2018) proposes a cost-sensitive classifier also in an attempt to model a specific loss function subject to fairness constraints. Results are evaluated empirically on a variety of data sets. Choi et al (2019) focuses on a specific family of classifiers, naive Bayes, and introduce the notion of a discrimination pattern alongside an algorithm for mining discrimination patterns in a naive Bayes classifier. The approach is iterative and seeks to eliminate such patterns until a fair model is obtained. An overview of the various notions of fairness can be found in Zemel et al (2013) and Dwork et al (2012). More comprehensive surveys are available in Friedler et al (2019) and Mehrabi et al (2019).

Problem Gambling Use Case

Data available

We work with two real-world data sets used to train the Random Forest harm prediction algorithms currently deployed by Playtech. The two gambling operators have different brands (one Bingo-focused and one Slot-Machine-focused), which will help demonstrate the diversity of circumstances even in a narrow ML domain.

The binary classification algorithms use data by players for whether they voluntarily self-excluded from the gambling platform during the analysis period, as an approximate proxy for experiencing harm. Only regular players who have been *live* in the platform for at least 1-2 months are included in the training data sets, given the focus of the algorithm on regular players. The open source Weka tool was used to replicate a comparable model to the deployed model with the same (approx. 40) behavioural input variables. A Random Forest model was trained on the raw, unbalanced training data, resulting in an accuracy under max-accuracy ROC curve for the two trained models within 1%pt of the deployed models. The two trained models (one for each operator) that are generated by this process are referred to in this paper as the baseline models.

These training data sets are enriched for the purpose of this study with gender data voluntarily supplied by players during the sign-up process. The gender variable can take three values: male (M), female (F) or unspecified/undeclared (U). Caveats remain with the quality of the available gender

[5] Reported by Kevin Peachey for BBC News on 18 November 2019, Sexist and biased? How credit firms make decisions https://www.bbc.co.uk/news/business-50432634

[6] 11 Nov 2019, Apple's sexist credit card investigated by US regulator. BBC News https://www.bbc.co.uk/news/business-50365609

[7] https://blogs.microsoft.com/ai/gender-skin-tone-facial-recognition-improvement/

[8] https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf

[9] https://ai.googleblog.com/2018/09/the-what-if-tool-code-free-probing-of.html

data, including possible gender bias in this voluntary supply of self-identification data as well as possible simplifications and distortions in having only two explicit categories for gender for players to select. Table 1 includes the summary descriptive data.

Bias definition

We identify three initial areas of analysis where metrics can usefully be analysed by gender: the gender balance in the training data, the self-exclusion rates in the training data, and the performance of the models for separate genders. All three sets of metrics are identified for reporting purposes, but only the model performance data is proposed as a metric for assessing potential problematic bias. Despite government-commissioned population surveys providing more detail by gender, population-wide surveys cannot be related to the gender ratios in an individual operator platform, as customer bases attracted by a particular brand are not representative of the overall gambling community.

Focusing on model performance, we are interested in a model that performs similarly well for each gender in terms of true positives (as a proxy for spotting those who are likely to be at risk) and true negatives (reducing any disruption or false alerts for players unlikely to be at risk). Since negative examples dominate in all populations and given the precautionary emphasis on identifying possible harm, the True Positive Rate (TPR) is the chosen primary performance metric for comparisons used in this paper. Given that we would not expect exact equality of performance by gender even in a perfectly fair algorithm, we also identify a tolerance threshold by which model performance might be identified as insufficient in that it should prompt action. For the exploratory purposes of this paper, we use a 2%pt difference in TPR performance among gender categories as such a threshold, noting that such a threshold must ultimately be informed by stakeholder consensus.

Experimental Results

Assessment of the influence of gender on the model (indirect discrimination)

Since gender is not included in the original algorithm there is no potential for direct use of gender in the model. However, gender may still be indirectly identified in the model via the correlations between other input variables and gender or other patterns in the data.

The standard initial investigation of relationships between variables is the correlation coefficient. For Operator 2, 8 of the 40 input variables have a correlation coefficient statistically significant at the 5% Bonferonni-adjusted level or better for male-reported gender, and 5 of the 40 input variables have the same correlation in the case of female-reported gender. However, this pattern is near-trivial by nature, in that the statistical significance reflects the large sample size rather than the meaningfulness of the co-variance. The r-squared from a linear regression using all statistically significant variables reveals that such variables only explain 0.6% of the linear variation in the male-reported gender dummy variable (RMSE of 0.30, RMSE across five-fold cross-validation varies from 0.30 to 0.31) and 1.1% of the linear variation in the female-reported gender dummy variable (RMSE of 0.48, RMSE of 0.48 in each of the five folds too). For Operator 1, there are no such statistically significant variables for the male-reported gender dummy and only two for the female-reported gender dummy, accounting for 2.2% of the linear variation (RMSE of 0.40, varying from 0.39 to 0.41 across five folds).

The correlation coefficient only identifies linear relationships, whereas the model in question is a Random Forest of depth 10 which is likely to identify non-linear patterns. While many common polynomial relationships in real-world gambling data might still be hinted at in a significant linear relationship, other relevant patterns would not. For instance, the Random Forest models used in this research were seen to identify relationships based on common values after the decimal point in a data set in which linear variation with gender had been removed by decomposition. Deducting the average value of a particular variable for each gender artificially results in zero linear correlation, but can result in such gender-driven patterns in the values after the decimal point, which remain exploitable by a Random Forest model.

Instead of linear correlation coefficients as a generic technique, we propose identifying the maximum possible level of indirect identification in a model-dependent manner, using a model with the same structure and parametrization as the original baseline model, an approach we call "model-matched indirect identification".

If the model were linear, with no interaction terms, bivariate linear correlations reflect the model structure and would be appropriate to capture possible indirect identification. In this case, we train a new model using the same ML method (Random Forests) with the same model parameter selection as the baseline model and the same set of predictor variables, but this time using gender as a target classification variable. The target variable from the baseline model, self-exclusion, does not appear in this new model.

The accuracy of this new model is seen as a bound on how well gender can be indirectly identified in the baseline model, since the new model is optimised to predict gender explicitly now, whereas the prediction of gender would only have been an indirect goal[10] of the baseline model, which is optimised to predict self-exclusion only. By using the same parametrization, we seek to find a maximum bound for the given use case (i.e. model + data) and to avoid the ambiguity of a possibly exponential variety of implicit interaction terms.

The gender-classification models produced by this approach have an out-of-bag (OOB) error[11] for Operator 1 of 0.5205 and for operator 2 of 0.4637. Collectively, this sug-

[10] Motivated only insofar as implicitly predicting gender midway through the model may later help to predict self-exclusion.

[11] Metric generated internally by Weka's Random Forest algorithm. This is an equivalent to a validation set performance measured for a fold from cross-validation, in that the RF algorithm deliberately excludes a set of observations in the construction of each tree. The OOB error measures the classification error rate for such excluded observations, taking the majority classification for each observation that has been excluded from various trees.

gests that there is little indirect identification of gender beyond a random guess based on the majority class.

Assessment of performance bias in baseline models (algorithmic fairness)

Table 1 reveals that male players outweigh female players on the slots-focused brand (1.6x prevalence) and male players are outweighed on the bingo-focused brand (3.5x prevalence). In both cases, undeclared gender is the most common group. Men tend to see higher levels of self-exclusion than women.

For Operator 1, there is little clear distinction in model performance by gender. The model is slightly better, based on TPR, at identifying women at risk than men, but stays within the 2%pt tolerance threshold. However, for Operator 2 there is a more marked higher model performance among female players than male players, with much higher TPR (+7.2%pts) and slightly higher overall accuracy (+0.8%pts). This gender delta by TPR is higher than the specified 2%pt threshold, prompting an exercise to see how it might be mitigated, as follows.

Options to enhance algorithmic fairness

In the online gambling use case, similar to other e-retail use cases, there is a strong sector preference for not compelling users to share sensitive data in order to use the services, both recognising the potential intrusiveness of such questions and the ease with which they can be inaccurately answered by those who would prefer not to be asked. For this reason, gender is a voluntary data point shared by players.

We test two mitigation methods for Operator 2 that do not require compulsory gender data: first, the inclusion of gender as an additional input variable (allowing Unspecified (U) to be one of its values). Secondly, we propose an ensemble method which is gender-blind at its deployment and which uses multiple gender-separated models in the ensemble aggregated to form an overall view on a player's risk. Naturally, if accurate gender data were assumed available for all players, other methods exist for reducing performance bias, provided stakeholders tolerate modelling structural differences by gender. For instance, separate classification thresholds could be set for men and women (potentially as part of gender-separated models) thus weighting false positives differently by gender, or output quotas could be set such that the top X highest-scoring male players and top Y highest-scoring female players are classified as at risk to meet a benchmark quota (potentially balanced against a decision rule that does not allow the quota to apply below a certain classification probability or clash with the above mentioned precautionary approach).

The first option above proved ineffective. Gender has little impact on the model. The original 0.1412 OOB error worsens marginally to 0.1440 with gender included. Male gender ranks 39 out of 42 input variables in terms of feature frequency in the Random Forest model, and female gender ranks 38 out of 42. The gender delta on TPR improves to 4.9%pts (reduced from 7.2%pts) but only with a worse TPR performance among women, with no improvement among men.

In the second option, blind-separate model, we train three separate models for confirmed male players, confirmed female players and gender-unspecified or undisclosed players. If any one model identifies a player as a likely self-excluder, the player is predicted to be at possible risk, reflecting the precautionary approach applied across many problem gambling identification strategies. As such, the overall classification approach is gendered but does not draw on gender as an explicit input variable once deployed. In this way, opt-in privacy of customers is preserved without a loss of access by customers to the best performing algorithms. A loss of access might happen if, for instance, one model were trained with gender data, while another (less accurate) model were trained without gender data, with the latter model used whenever a customer chooses not to share gender data. In our approach, gender data is only required for a sample of the players, which might be developed from voluntarily provided data (as done here) or via an ad-hoc collection for the sole purpose of such a model. This approach reduces the gender disparity in TPR, but at the cost of the true negative rate (TNR) and accuracy. Male TPR increases from 46.5% to 54.7% and reduces the gender delta from 7.2%pts to 4.0%pts. It is important to note that this improvement in TPR and reduction in delta cannot be achieved by simply altering the classification thresholds in the baseline model: the delta increases to 7.3%pts in the baseline model if its classification threshold is adjusted until the male TPR matches the male TPR from the blind-separate model. This provides confidence that the blind-separate model, in its use of gender-insights, is providing additional value in the identification of players at possible harm.

Nonetheless, as mentioned, this reduction in gender disparity by TPR comes at the cost of accuracy and TNR. The OOB error is higher for men, which has the smaller sample of the two confirmed genders (0.1452 vs 0.1361 for women). TNR decreases from 96.7% to 95.3% for women, from 98.1% to 91.9% for men and from 97.2% to 94.7% for unspecified gender.

This may be an acceptable loss of performance in exchange for reduced gender disparity, given the gambling industry focus on the precautionary principle, but would require exploration with sector stakeholders. It is also possible that a larger exercise may result in model choices that entail other forms of compromise: one disadvantage of the blind-separate model is that it reduces the sample size available for training in each gender group. Improvements might be expected with larger training data sets and the application of data set balancing techniques (in Playtech's deployed algorithms, the SMOTE technique is used to generate balanced data and it is not used here; see Percy et al (2016) for details).

Lessons Learned, Conclusion and Future Work

The purpose of this paper has been to report work by Playtech, a provider of B2B and B2C gambling services, to investigate the role of gender in its gambling harm identification algorithms. We identify five key lessons learned to

Possible metric by gender (F/M)	Operator 1 (slots-focused brand, n = 4,340)	Operator 2 (Bingo-focused brand, n=18,275)
Gender balance in training data	F: 20.6% M: 32.6% U: 46.8%	F: 36.5% M: 10.4% U: 53.1%
Self-exclusion outcome	F: 20.4% M: 24.4% U: 16.8%	F: 17.1% M: 18.7% U: 22.3%
Baseline RF model TPR	F: 67.0% M: 65.3% U: 66.5%	F: 53.7% M: 46.5% U: 52.9%
Baseline RF model TNR	F: 94.4% M: 95.1% U: 95.0%	F: 96.7% M: 98.1% U: 97.2%
Baseline RF model accuracy	F: 88.8% M: 87.9% U: 90.2%	F: 89.3% M: 88.5% U: 87.4%

Table 1: Gender metrics from two gambling operators.

date as part of an ongoing project to improve practice:

- The diversity of ML use cases, data sets and stakeholder priorities is such that there is no single stance on what algorithmic fairness should be prioritised or how it should be enhanced. For the two models in the gambling harm identification use cases, we have found negligible levels of indirect gender identification in gender blind models. Focusing on gender disparities in true positive rates, we found a meaningful disparity in one model but not the other. The same technique that reduced gender disparity on the target model would have increased it on the other model in the other data set, so we should not assume consistency from one context to another.
- Analysis of bias requires investing resources in the definition and defence of unbiased benchmarks and the specification of a tolerance threshold. Since bias can exist either above or below any given benchmark, random variation makes it impossible to achieve an exact ongoing fit. The margin of error which can be tolerated depends on what stakeholders find material, worth the apportion of resources and the level of variation in the values of a protected variable as measured over time and over different data set samples.
- Exercises to improve algorithmic fairness need to be incorporated into overall business priorities, most likely engaging appropriately balanced stakeholder groups, rather than treated as a separable analytical exercise. This is both because judgement calls need to be made as part of the exercise and because adjusting practice based on insight may require the balancing of multiple objectives, some of which may be competing objectives.
- Indirect discrimination needs to be analysed as a feature of a specific model rather than a feature of the data set. For instance, a target variable such as gender may be mapped in diverse ways against other variables in the data set depending on the complexity of the model (e.g. linear, polynomial, interaction-dependent, integer/decimal structure, etc). Indirect discrimination is driven by whether your model can exploit a particular pattern, rather than by other patterns that might exist.
- Any analysis of fairness is inevitably limited, both because of changing expectations and the potential breadth of the topic. As such, it is important to treat it as a process rather than a one-off exercise and to recognise the limits in any one exercise. In this initial exploratory work, for instance, it is unclear what biases a voluntary provision of gender data might introduce. Gender bias is also likely to exist elsewhere in the technical and cultural institutions surrounding gambling, the self-identification of problem gambling, and the socio-economic system on which gambling is embedded; it is unclear how such biases might influence training data and the resulting AI algorithms. More specifically, this exploratory analysis has focused on a sample of regular players and two operators, and it may not be reflective of early-stage players or players with other operators.

In what concerns future work, in the gambling sector, our next step is convening a working group to apply this at a larger scale and discussing compromises among competing objectives. Such a group might comprise domain experts (e.g. ML experts and data scientists, legal counsel, experts in the target variable, experts in the use case), managers and external representatives who provide challenge and validity as part of the overall exercise, ensuring representation of individuals from different groups in the target socio-demographic variables. In doing so, the objective is to improve safer gambling outcomes across all cohorts and the scope can be expanded to include the design and evaluation of industry level interventions as well as risk identification algorithms.

On the ethics of AI more generally, we shall develop a general framework out of our approach to investigating algorithmic fairness in other use cases in the sector, supported by a taxonomy of the diverse techniques available to improve fairness. We invite comment, engagement and challenge on this paper as part of the broader project to improve practice and to develop relevant and industry-specific AI principles.

References

Advisory Committee on Equal Opportunities for Women and Men for the European Commission. (2020). Opinion on Artificial Intelligence - opportunities and challenges for gender equality (published 18 March 2020).

Agarwal, Alekh; Beygelzimer, Aliiia; Dudfk, Miroslav; Langford, John and Hanna, Wallach. A Reductions Approach to Fair Classification, 35th International Conference on Machine Learning, ICML 2018, Stockholm, Sweden, July 2018.

Baggio, S., Gainsbury, S., Starcevic, V., Richard, J., Beck, F., Billieux, J. (2018). Gender differences in gambling preferences and problem gambling: a network-level analysis, International Gambling Studies, 18:3, 512-525.

Bolukbasi, T., Chang, K., Zou, J., Saligrama, V., Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. Available via arXiv:1607.06520v1 [cs.CL] 21 Jul 2016.

CDEI. (2020). AI Barometer Report: June 2020. London: Centre for Data Ethics and Innovation, UK.

Choi, YooJung; Farnadi, Golnoosh; Babaki, Behrouz and Broeck, Guy Van den. Learning Fair Naive Bayes Classifiers by Discovering and Eliminating Discrimination Patterns. In Proc. AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, February 2020.

Dragicevic, S., Garcez, A., Percy, C., Sarkar, S. (2019). Understanding the Risk Profile of Gambling Behaviour through Machine Learning Predictive Modelling and Explanation. KR2ML 2019, Workshop at 33rd NeurIPS Conference, Vancouver, Canada, December 2019 (available via https://kr2ml.github.io/2019/papers/).

Dwork, Cynthia; Hardt, Moritz; Pitassi, Toniann; Reingold, Omer and Zemel, Richard. Fairness through awareness, Innovations in Theoretical Computer Science Conference, ITCS2012, MIT CSAIL, Cambridge MA, January 2012.

Friedler, Sorelle A; Choudhary, Sonam; Scheidegger, Carlos; Hamilton, Evan P; Venkatasubramanian, Suresh and Roth, Derek. A Comparative Study of Fairness-Enhancing Interventions in Machine Learning, In Proc. 2019 ACM Conference on Fairness, Accountability and Transparency, Atlanta, GA, January 2019.

Mehrabi, Ninareh; Morstatter, Fred; Saxena, Nripsuta; Lerman, Kristina and Galstyan, Aram. A Survey on Bias and Fairness in Machine Learning, KR2ML Workshop at NeurIPS'19 Conference, Vancouver, Canada, December 2019 (available via https://kr2ml.github.io/2019/papers/).

Gambling Commission (2018). Participation in gambling and rates of problem gambling – England 2016: Statistical report. Birmingham, GC, UK.

Goodman, B. (2016). A Step Towards Accountable Algorithms? Algorithmic Discrimination and the European Union General Data Protection. 29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 2016.

Kleinberg, J., Ludwig, J., Mullainathan, S., Rambachan, A. (2018). Advances in big data research in economics: Algorithmic fairness. AEA Papers and Proceedings 2018, 108: 22–27 https://doi.org/10.1257/pandp.20181018, 2018.

Lundberg, S., Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, December 2017.

McCarthy, S., Thomas, S.L., Bellringer, M.E. et al. (2019). Women and gambling-related harm: a narrative literature review and implications for research, policy, and practice. BMC Harm Reduction Journal, 16-18 2019.

Moerel, L. (2018). Algorithms can reduce discrimination, but only with proper data. Publ. 16 Nov 2018 by IAPP, 2018.

Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science 25 Oct 2019: 447-453. https://science.sciencemag.org/content/366/6464/447.

Percy, C., França, M., Dragičević, S., Garcez, A. (2016): Predicting online gambling self-exclusion: an analysis of the performance of supervised machine learning models, International Gambling Studies, 2016.

Sarkar, S., Weyde, T., Garcez, A., Slabaugh, G., Dragicevic, S., Percy, C. (2016). Accuracy and interpretability trade-offs in machine learning applied to safer gambling. CEUR Workshop Proceedings, 1773. Dec. Available via http://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper10.pdf.

Stenstrom, E., Saad, G. (2011). Testosterone, financial risk-taking, and pathological gambling. Journal of Neuroscience, Psychology, and Economics, 4(4), 254–266.

Su, W., Han, X., Yu, H., Wu, Y., Potenza, M. (2020). Do men become addicted to internet gaming and women to social media? A meta-analysis examining gender-related differences in specific internet addiction. Computers in Human Behavior, Volume 113, 2020.

Suresh, H., Guttag, J. (2020). A Framework for Understanding Unintended Consequences of Machine Learning. Available via arXiv:1901.10002v3 [cs.LG], 2020.

Venne, D., Mazar, A., Volberg, R. (2019). Gender and Gambling Behaviors: A Comprehensive Analysis of (Dis)Similarities. Int J Ment Health Addiction, 2019.

White, A., Garcez, A (2020). Measurable Counterfactual Local Explanations for Any Classifier. In Proc. 24th European Conference on Artificial Intelligence, ECAI 2020, Santiago de Compostela, Spain, Aug 2020.

Wong, G., Zane, N., Saw, A., Chan, A. K. (2013). Examining gender differences for gambling engagement and gambling problems among emerging adults. Journal of gambling studies, 29(2), 171–189.

Zemel, Richard; Wu, Yu; Swersky, Kevin; Pitassi, Toniann and Dwork, Cynthia. Learning Fair Representations, 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, June 2013.