Journal of Gambling Issues

CDS PRESS

Open Access **Original Research**

Online Gamblers' Preferences for Performance and Fairness in Artificial Intelligence Systems for Gambling Harm Detection

W. Spencer Murch[1,3], Martin French[2.4], Sylvia Kairouz[2.5]

[1]Department of Psychology, University of Calgary, Canada
[2]Department of Sociology and Anthropology, Concordia University, Canada
[3]ORCiD: 0000-0003-2780-3578
[4]ORCiD: 0000-0001-8724-5139
[4]ORCiD: 0000-0002-8788-4456
*Corresponding author: Spencer Murch: spencer.murch@ucalgary.ca

Abstract. Amid the rising popularity of online gambling websites and apps, researchers have sought to develop Artificial Intelligence (AI) systems that identify users at-risk for gambling-related harms. However, existing systems' classification performance is imperfect, and even strongly-performing systems may see their performance drift over time or vary across sociodemographic groups in unfair ways. Taking inspiration from patient-centered healthcare, we investigated the preferences that online gamblers have for the performance and fairness of automated gambling harm-detection systems. Canadian online gamblers ($N = 107$) completed a data visualization task which asked them to judge the classification performance, fairness, and real-world readiness of 36 hypothetical detection tools that depicted differing classification performances for people experiencing and not experiencing online gambling problems. Multilevel regression models revealed that more sensitive (b = 0.03, p < .001) and more specific ($b = 0.01$, $p < .001$) AI systems were rated as better-performing. Presented systems were rated as significantly less fair when they depicted poorer classification performance for older (b = -0.81, p < .001) or younger (b = -0.69, p < .001) persons. Participant-defined performance standards for systems to be 'real-world-ready' suggested a minimum classification sensitivity at 73.71% and specificity at 69.39%. These results suggest that Canadian online gamblers have sensible and realistic desires for the performance of automated gambling harm reduction systems. However, further improvements to many existing AI systems are needed before end users may consider this technology ready for real-world deployment.

Keywords: Gambling, Artificial intelligence, Fairness, Machine learning, Prevention, Data visualization.

Citation: Murch, W.S., French, M., Kairouz, S. (2025). Online Gamblers' Preferences for Performance and Fairness in Artificial Intelligence Systems for Gambling Harm Detection. Journal of Gambling Issues.

Editor-in-Chief: Nigel Turner, Ph.D.

Received: 10/10/2024
Accepted: 09/03/2025
Published:09/27/2025

Introduction

There is an old joke that goes like this: "What picks three apples, belches smoke, and explodes? A machine designed to pick four apples." The point is that machines do not always accomplish everything they were designed to do. Even the rare case of a perfectly-performing machine is defeated over time as entropy guarantees all things must eventually break down. While some machines' purposes are trivial, others affect human lives in profound or irreversible ways. How, then, should we decide when it is acceptable to use important-but-imperfect machines that could have serious consequences for people's lives?

Answering this question within the field of gambling research is increasingly important as gambling behaviours and gambling-related harms continue to evolve. Over the last two decades, several gambling nations have seen an ongoing migration of traditionally land-based gambling activities towards online variants accessed via the internet. Beginning around 2003 with a boom in internet poker gambling, buoyed by regulatory changes that enabled fantasy and other online sports wagering, and accelerating in 2020 with COVID-19-related closures of physical gambling venues, more and more people participate in gambling via the internet (Das, 2021; Shead et al., 2008; Stark & Robinson, 2021). This shift towards online gambling is concerning both because existing harm reduction or 'responsible gambling' strategies were overwhelmingly designed and validated in physical gambling venues (Ladouceur et al., 2017; Reynolds et al., 2020), and because a growing body of evidence suggests online gambling activities – being covertly and constantly accessible – are more-closely linked to gambling-related problems than many other formats (Gainsbury, 2015; Gooding & Williams, 2023; Olason et al., 2011; Petry, 2006; Williams et al., 2015).

To begin addressing the rising tide of online gambling-related harms, gambling researchers increasingly look to Artificial Intelligence (AI) technologies, with a particular emphasis on the subfield of machine learning (Ghaharian et al., 2022). Whereas traditional statistical approaches seek to describe scientific observations in a parsimonious and transparent manner, machine learning-based approaches instead tend to elaborate (often) more complex and (often) less transparent statistical models with the aim of achieving better predictive performance in future samples (Badillo et al., 2020). In practice, this means creating systems that ingest large volumes of online gambling transaction data (obtained from a gambling operator, financial service provider, etc.) in order to generate predictions that distinguish online gamblers who are at elevated risk for harm from those who are not (Braverman & Shaffer, 2012). This approach thus promises to enable earlier detection and improved treatment outcomes via interventions tailored to the predicted needs of those being reached.

Several studies have shown that AI models can ingest transactional data from online gambling websites and make accurate predictions about harm-relevant variables such as voluntarily excluding one's own access to gambling (Akhter, 2017; Finkenwirth et al., 2020; Haeusler, 2016; Percy, 2020; Percy et al., 2016; Xuan & Shaffer, 2009), account closure (Braverman & Shaffer, 2012; Philander, 2014), negative contacts with customer support staff (Haefeli et al., 2015), and responses on self-report problem gambling questionnaires (LaPlante et al., 2014). Advancing the latter, recent research has also explored classification models that predict scores on the Problem Gambling Severity Index (PGSI; Ferris & Wynne, 2001), a prominent self-report instrument that assesses past-year gambling problems (Auer & Griffiths, 2022; Kairouz et al., 2023; Luquiens et al., 2016; Murch et al., 2023, 2024b). The key advantage to this approach is that the PGSI comes from end users' self-appraisals rather than 'objective' metrics like voluntary self-exclusion that may capture only small portions of those experiencing harm (a key critique of earlier work on models predicting voluntary self-exclusion; Finkenwirth et al., 2020). Models geared towards predicting problem gambling risk categories on the PGSI have consistently demonstrated a capacity for strong classification performance at the time of their initial validation (Auer & Griffiths, 2022; Kairouz et al., 2023; Luquiens et al., 2016; Murch et al., 2023).

That said, no existing system has demonstrated perfect classification performance at the time of its initial validation, and it has been established that AI models involving humans see their classification performance 'drift', worsening over time as the population evolves (Lu et al., 2019). Going further, the emerging field of *Algorithmic Fairness* stresses the importance of considering how AI systems can impact people from different sociodemographic backgrounds (Benjamin, 2019; Corbett-Davies & Goel, 2018; Sweeney, 2013). We previously showed (Murch et al., 2024a) that AI models designed to detect at-risk online gamblers based on their financial transactions can be undermined by established dependencies between the rate of gambling-related problems and the age group to which one belongs (Potenza et al., 2019). Percy and colleagues (2020) promisingly provided an ensemble-modelling approach that diminished (but did not completely erase) sex-related disparities in AI-based predictions made about users of two online gambling platforms.

It is thus clear that there are several ways automated tools – including AI systems – may break down when trying to detect at-risk online gamblers for individualized treatment and support. The promise of this technology is thus at odds with the practical realities of its application. Nevertheless, if an automated detection tool could be identified as 'good enough' at its classification task, 'fair enough' across sociodemographic groups, and 'ready enough' for real world use, then AI technology may yet be suitable for the task of online gambling harm reduction. Consistent with the public health principle of patient-centered care (Stewart et al., 2000), we argue that it should be online gamblers themselves who ultimately decide where these lines are drawn. Going further, it is valuable at this stage to recognize that online gamblers are at once: (1) knowledgeable people with a direct interest in the final design of AI systems that may affect them personally, (2) the group most-comprised of people experiencing online gambling-related harms, and (3) the ultimate decision-makers as to whether a harm prevention system will be utilised (if a system is effective but strongly undesirable, it may drive users away from a platform).

We similarly recognize that differences in individuals' identities and experiences may influence how they perceive these systems to function, and which groups they believe may be more or less deserving of strong classification performance. For example, it is conceivable that individuals at lower risk for gambling problems may place relatively less importance on correctly identifying at-risk individuals, or that individuals who perceive themselves as 'younger' may place relatively greater value on maximizing the correct classification of younger people (even if that means allowing unfairness between age groups). For this reason, we additionally modeled individuals' own age and problem gambling status. We believe this is the first study to ask online gamblers about their views on the kind of AI models proposed over the last 10 years.

The current study

To investigate perceptions of AI models for online gambling harm detection, we recruited Canadian online gamblers and asked them to judge different data visualizations depicting hypothetical tools for automatically detecting at-risk online gamblers. These visualizations were manipulated to depict differences in classification performance separately for 'at-risk' and 'not-at-risk' gamblers, as well as between two age groups ('older' and 'younger') within the at-risk and not-at-risk categories (Figure 1). For each visualization, participants rated whether the system seemed well-performing, fair, and suitable for real-world use.

We hypothesized that:

H1: Online gamblers will report differing degrees of perceived efficacy for automated detection tools that depict high proportions of false positive versus false negative cases. Specifically:
H1.1: Information pertaining to the systems' overall and age-specific classification performance will impact online gamblers' ratings of the models' classification performance, fairness, and real-world readiness.
H1.2: Stimuli depicting discrepancies in age-specific classification performance will be rated as less fair than those showing no discrepancy.
H1.3: Participants' own age, and self-reported levels of gambling-related problems will significantly impact ratings of perceived efficacy and fairness.

In discussing the study's results, we draw comparisons against previously-published machine learning models that – in various ways – sought to detect at-risk gamblers for the purposes of harm prevention. For each, relevant performance metrics are summarized in order to determine if any existing system has achieved agreeable performance standards outlined by this study's participants. These comparisons thus evaluate whether automated gambling harm detection technologies currently perform sufficiently well for real-world use, as judged by people from the population (online gamblers) that stands to be most impacted by such technologies' (mis)behaviours.

Methods

Participants

Canadian online gamblers aged 18 years or older were recruited via prolific.com (Prolific, United Kingdom), a participant recruitment platform with a large userbase and widespread use in the social sciences. A three-question pre-screen survey determined the eligibility of Prolific users based in Canada by confirming their: (1) province of residence, (2) date of birth, and (3) participation in at least one online gambling activity during the prior 12 months. *A priori* power calculations for the study's multilevel regression models indicated a necessary sample size of $N = 109$ participants in order to achieve 80% power given a moderate effect size ($f^2 = 0.15$), 8 predictor variables (the final design of the data visualization task ultimately dropped two factors, leaving 6 predictor variables per model), and a conventional α-level for statistical significance (.05). A total of 286 Prolific users were screened before 109 eligible participants completed the full task. Two participants were excluded from analysis after failing an attention check question embedded within the survey, making the final sample size 107.

All respondents and participants gave informed consent consistent with the Declaration of Helsinki and the protocol approved by the host institution's Office for Research Ethics (#30018600). Participants were told that the study could take up to 5 minutes for pre-screening and up to 60 minutes for full participation. Pre-screen respondents were paid $1.70 CAD for a median 1 minute and 12 seconds of their time (min = 0m35s, IQR = 0m55s – 1m47s, max = 130m35s). Task participants were paid $20.29 CAD for a median 17 minutes and 35 seconds of their time (min = 6m3s, IQR = 13m5s – 25m46s, max = 92m14s).

Questionnaires

All surveys and tasks were offered in English and French. In addition to providing their age and province of residence in the pre-screen survey, participants were asked their biological sex (female / male / open-ended response) and gender identity (man / woman / non-binary / open-ended response). This sociodemographic information was used to evaluate the similarity of this participant sample to the Canadian population and the subpopulation of Canadian online gamblers.

After providing sociodemographic information, participants were asked to provide additional details pertaining to their participation in different gambling activities. Specifically, they were asked whether – during the prior 12 months – they had participated in gambling via: (1)

lottery or raffle tickets, excluding sports lottery tickets, (2) instant lottery tickets, such as scratch, break-open or pull-tabs, or instant online games, (3) electronic gambling machines, such as slot machines, VLTs, electronic blackjack, electronic roulette or video poker, (4) casino table games such as poker, blackjack, baccarat, or roulette, (5) sports betting activities such as hockey, football, horseracing, billiards or golf including pools, sports lottery, and bets made with friends, (6) bingo games, excluding instant bingo games, or (7) financial matters such as day trading, penny stocks, shorting, options, or currency futures. These categories were drawn from the gambling participation module included in the 2018 Canadian Community Health Survey (Statistics Canada, 2019). Participants were additionally asked to rate the percentage of their past-year gambling that occurred online.

Finally, participants completed the Problem Gambling Severity Index (PGSI), a popular and widely-validated measure of past-year gambling problems (Currie et al., 2013; Dellis et al., 2014; Ferris & Wynne, 2001; Holtgraves, 2009; Miller et al., 2013). The PGSI employs a hybrid conceptualization of gambling problems, concurrently measuring specific problematic behaviours (e.g., borrowing money to gamble), addiction-relevant dimensions of problematic gambling (e.g., loss of control over betting; loss chasing), and health-relevant outcomes related to problematic gambling (e.g., stress or anxiety). Each item is given a frequency rating ranging from 0 (“never”) to 3 (“almost always”), and the scale’s nine items are summed for a total score out of 27. Widely-used interpretive categories for the PGSI define Non-problem gambling (scores of 0), Low-risk (1-2), Moderate-risk (3-7), and Problem gambling (8+) groups (but see Currie et al., 2013; Williams & Volberg, 2014; Samuelsson et al., 2019).

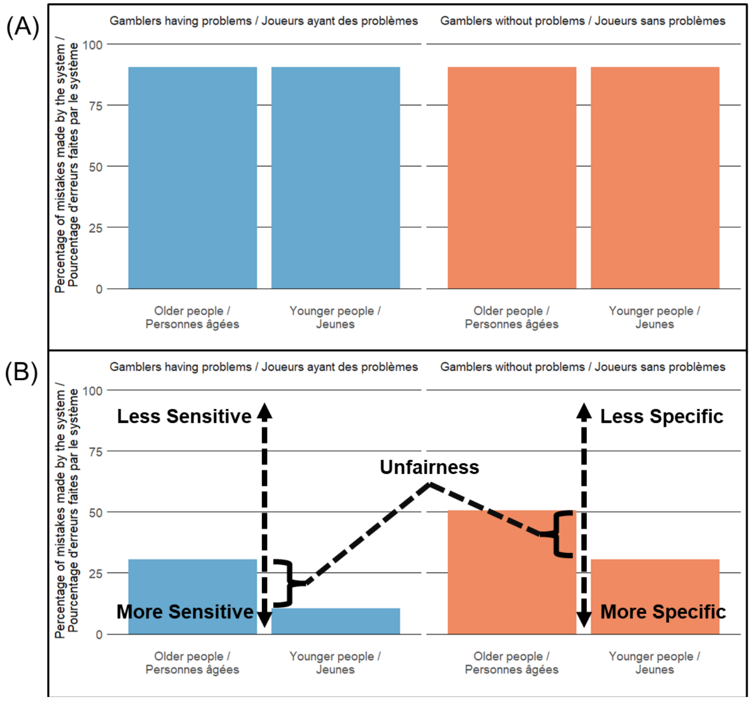
Data visualization task

We employed a data visualization task which asked participants to evaluate 36 hypothetical ‘tools’ for the detection of online gambling problems. Asking participants to judge stimuli that had been systematically manipulated enabled us to see how differences in these tools’ depicted functioning impacted participants’ ratings of their performance, fairness, and readiness for real-world use. The design and implementation of this task was evaluated using a standardized checklist for conjoint analysis designs (see Supplementary Materials; Bridges et al., 2011).

In this task, participants were asked to “imagine that scientists have developed a tool that analyzes online gambling websites to identify which people are experiencing gambling problems.” They were then told that “[the] tool could allow counselors to more-quickly help people who are having problems.” They were informed that this hypothetical system is not perfect, and might in some cases: (1) fail to notice when someone is having problems, (2) think someone has gambling problems when they actually do not, or (3) perform in ways that are not equally good for older and younger people.

Participants were then asked to view 36 data visualizations depicting hypothetical classification performance levels for one version of the system (see Figure 1 for examples; see also van Berkel et al., 2021), and provide judgments relating to several aspects of each one. Drawing inspiration from earlier visualization work (van Berkel et al., 2021), the authors drafted and iteratively simplified a bar chart-style visualization capable of depicting three distinct manipulation conditions (see below). These performance levels were depicted for each stimulus on two bar graphs presented side-by-side. The y-axis of these graphs reported to show the percentage of mistakes made by the system, ranging from 0% to 100%. Incorrect classifications were displayed to avoid inducing confusion as task instructions explained the systems' mistakes rather than their successes. Percentage values were displayed to simplify judgments since no information were provided for either positive or negative predictive value (which would be present if frequencies were used instead of percentages). This removes any implication of the prevalence of gambling problems, instead presenting people experiencing and not experiencing gambling problems merely as independent groups. The full range of possible percentages was used to avoid implying how the system *ought* to perform in one way or another.

Figure 1. Example stimuli



Note: Two example stimuli among the 36 total visualizations judged by each participant. (A) depicts an example AI system that has both poor sensitivity (tall blue bars) and poor specificity (tall orange bars) rates, but which performs in a manner that is fair between age groups. (B) depicts a system that appears reasonably sensitive and specific, but which performs noticeably better among younger people. The three stimulus manipulations (sensitivity/FNR, specificity/FPR, and fairness) are annotated on panel B. All text was shown in both English and French.

Displayed bars were then stratified based on whether individuals were purported to actually be experiencing gambling problems, and whether individuals belonged to 'younger' or 'older' age groups. Age was selected for this "model fairness" manipulation because the experience of gambling problems covaries with individuals' age (Murch et al., 2024a; Potenza et al., 2019), making it a realistic candidate for real fairness evaluations in the development of AI-based gambling harm detection systems.

Given this structure, each data visualization thus separately displayed a hypothetical model's False Negative Rates (FNR, defined as the percentage of people with gambling problems that the model failed to flag as at-risk; 1 - sensitivity) and False Positive Rates (FPR, defined as the percentage of people without gambling problems incorrectly flagged as at-risk; 1 - specificity) for 'older' and 'younger' people. In other words, each stimulus was comprised of three manipulated variables: FNR, FPR, and fairness condition. The FNR and FPR manipulations for the task (whose results are reported in terms of their inverse values, sensitivity and specificity) were each varied and counterbalanced across six percentage values (10%, 20%, 40%, 60%, 80%, and 90%) that were then adjusted upward or downward by 10% depending on the fairness manipulation (fair [no adjustment], favours younger [-10% to younger FNR and FPR, +10% to older FNR and FPR,], and favours older [+10% to younger FNR and FPR, -10% to older FNR and FPR]), which was Latin-square balanced to prevent systematic differences in depicted classification performance between fairness conditions. Stimuli were presented in random order for each participant. The minimum and maximum values of 10% and 90% were set to prevent impossible percentage values (i.e., <0% or >100%) from being inadvertently displayed while keeping the magnitude of the fairness manipulation consistent throughout.

For each stimulus, participants were asked to indicate their level of agreement with three statements: "based on the information provided, I would say the system works well for online gamblers" (hereafter "goodness"), "Based on the information provided, I would say that the predictions this system makes are fair" (hereafter "fairness"), and "I would say that this system is good enough to be used in the real world" (hereafter "readiness"). Each judgment was provided on a 7-point Likert scale that was integer scaled with equal spacing ("Strongly disagree," "Disagree," "Somewhat disagree," "Neither agree nor disagree," "Somewhat agree," "Agree," "Strongly agree"). To aid in participants' judgments, they were provided several 'hints' on screen, informing them that "tall bars mean lots of mistakes for people with (blue) or without (orange) gambling problems," and "different bar heights with the same colour indicate a different number of mistakes for older and younger people." Full task instructions and all stimuli are archived (**Note 1**).

Analysis

All analyses were completed in R version 4.3.2 using the 'nlme', 'MuMIn', 'ggplot2', and 'rColorBrewer' packages (Bartoń, 2024;

Neuwirth, 2014; Pinheiro et al., 2019; R Core Team, 2023; Wickham, 2016). All scripts and anonymized data have been archived (**Note 1**). For each of the study's three outcome variables (goodness, fairness, and readiness), a repeated-measures model was fit that nested individual traits and responses within each participant. Specifically, multilevel models (see Field et al., 2012 p. 873) were fit which nested several hypothesis-defined level 1 variables (see below) within a single level 2 variable (Participant ID) that allowed model intercepts to vary within each participant. Level 1 variables included the visualizations' depicted sensitivity (1-FNR) and specificity (1-FPR) conditions (each centered on 0%), the depicted fairness condition (as a factor with reference category "fair"), participants' self-reported PGSI score (centered on zero), participants' reported age group (as a factor with reference category "30-49"), and a participant age-by-depicted fairness interaction term. The need for multilevel modelling was verified by testing a random intercepts-only model against the null model (see archived analyses). The appropriateness of model fits was verified via visual inspection of their standardized residual and QQ plots. For each final model, pseudo coefficients of determination (R^2) were computed for both the Level 1 (fixed) and Level 2 (participant) effects.

For each of the three models (goodness, fairness, and readiness), predicted scores were used to compute Root Mean Squared Error (RMSE) statistics in order to indicate the mean prediction error. Based on these fits and RMSE values, subsets of 'minimally-agreeable' cases were defined as those with predicted values greater than a response of "somewhat agree" plus one RMSE. In effect, these subsets identified cases where the predicted agreeability for goodness, fairness, or readiness, would need to be *more-wrong than average* in order for that case to be less than *at-least somewhat agreeable*. The characteristics of these subsets were summarized to identify necessary model performance conditions and respondent characteristics for minimally-agreeable systems.

Finally, we scanned recent publications and existing reviews of data science applications to gambling research (Ghaharian et al., 2022) to identify studies that previously attempted to develop automated gambling harm detection systems. The performance of these earlier reports was compared against minimally-acceptable thresholds for real-world readiness to determine if AI technology is currently fit for the purpose of gambling harm detection (Figure 4).

Results

The participant sample reasonably reflected Canada's English-speaking provinces, but few people residing in Quebec and no people residing in Canada's three northern territories were recruited (Table 1). Of all participants (*N* = 107), only 5 reported residing in Québec, and only two participants elected to complete the study in French. Additionally, most participants reported being cisgender men, but cisgender women were also well-represented in the sample (Table 1). Participants aged 30 to 49 years made up the largest study group (*n* = 60), but those aged 18 – 29 years (*n* = 32) and those 55+ (*n* = 15) were also well represented.

Nevertheless, the online gamblers who made up this sample reported a broad range of gambling activities over the past year (Table 1). Most reported having engaged in lottery and instant lottery (‘scratch card’) gambling (Table 1). Many participants also reported sports wagering, electronic gaming machine use, speculative financial trading, bingo, and gambling on table games at casinos. Further, participants were not just experienced online gamblers, but rather *primarily* online gamblers; most participants (*n* = 56) reported doing at least 80% of their gambling online. Participants reported a wide range of past-year gambling problems, with most (*n* = 59) occupying either the Moderate-risk or Problem gambling categories of the PGSI. Fewer than one in five (*n* = 18) participants reported zero past year gambling problems. Thus, this sample was composed of online gamblers who arguably have a personal interest in seeing improvements made to existing online gambling harm prevention techniques.

Table 1. Sociodemographic traits and gambling involvement among participants

		Frequency
Language		
	English	105
	French	2
Age group		
	18 – 29	32
	30 – 49	60
	50+	15
Province of residence		
	Atlantic	9
	British Columbia	11
	Ontario	61
	Prairie	21
	Quebec	5
	Territories	0
Sex		
	Female	42
	Male	64
	Prefer not to say	1
Gender identity		
	Man	64
	Nonbinary	1
	Woman	40
	Prefer not to say	2
PGSI risk category		
	Non-problem gambler (0)	18
	Low-risk (1 – 2)	30
	Moderate-risk (3 – 7)	45
	Problem gambler (8+)	14
Gambling participation		
	Lottery	79
	Instant lottery / scratch cards	69
	Sports wagering	50
	Electronic gambling machines	48
	Speculative financial trading	34
	Casino table games	29
	Bingo	12
	Other	5
Online gambling participation (% of all gambling)		
	0% – 19%	12
	20% – 39%	9
	40% – 59%	10
	60% – 79%	19
	80% – 100%	56
	Prefer not to say	1

Note: Atlantic and prairie provinces were aggregated to reduce participant re-identifiability after the fact.

Data visualization task results

Relationships between dependent variables

This study's dependent variables were highly collinear. Averaging ratings for each outcome variable within participants, significant positive correlations were found between whole-task ratings of real-world readiness and both goodness ($r(105) = .96$, $p < .001$) and fairness ($r(105) = .87$, $p < .001$). Overall goodness and fairness ratings, in turn, were positively correlated ($r(105) = .83$, $p < .001$). Averaging participant responses for each outcome variable within the 36 stimuli, real-world readiness ratings were again positively related to goodness ratings ($r(34) = .99$, $p < .001$) and fairness ratings ($r(34) = .94$, $p < .001$), which were in turn positively related to each other ($r(34) = .92$, $p < .001$). Although ratings were mostly consistent, there are several stimuli where mean ratings noticeably diverge. Mainly, these divergences concerned stimuli that were poor in terms of classification performance, but were nevertheless fair, and those that were strong overall classifiers, but depicted some degree of unfairness.

Ratings of system performance (goodness)

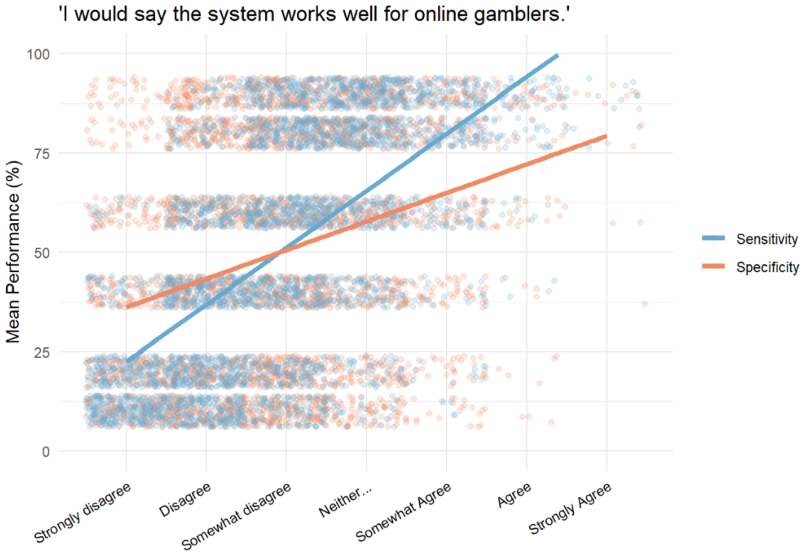
Participants' ratings of whether depicted detection systems appeared to perform well in detecting online gamblers experiencing (and not experiencing) problems depended on the experiment's sensitivity, specificity, and fairness manipulations. Model fit was strong overall ($R^2_{\text{Total}} = 0.49$; Table 2A). For sensitivity and specificity, these relationships were positive, with goodness ratings increasing as sensitivity ($b = 0.03$, $p < .001$) and specificity ($b = 0.01$, $p < .001$) were shown to increase. Notably, the coefficient relating to sensitivity shows a significantly steeper slope (Table 2) than the coefficient relating to specificity, indicating that participants' goodness ratings were more closely related to how well these systems detected people experiencing gambling problems (Figure 2). Goodness ratings were negatively related to the experiment's fairness condition such that systems favouring older ($b = -0.23$, $p < .001$) or younger ($b = -0.40$, $p < .001$) people were both rated as performing significantly worse than systems that performed equally across age groups. These effects differed significantly from one another such that systems favouring younger people (and thus disadvantaging older people) were rated as performing even worse than those whose unfairness favoured older people. Goodness ratings did not significantly depend on participants' PGSI scores, their own age, or the interaction between their age group and the depicted fairness condition.

Table 2. Modelling results for the data visualization task

	Model					
	A: "Goodness"		B: "Fairness"		C: "Real-world readiness"	
Level 2	**L ratio**	***p***	**L ratio**	***p***	**L ratio**	***p***
Intercept (Participant ID)	755.88	< .001	886.37	< .001	916.93	< .001
Slope (None)	-		-		-	
	$\mathbf{R^2_{Level\ 2}}$	0.23	$\mathbf{R^2_{Level\ 2}}$	0.25	$\mathbf{R^2_{Level\ 2}}$	0.27
Level 1	***b* [95% CI]**	***p***	***b* [95% CI]**	***p***	***b* [95% CI]**	***p***
Intercept	0.03 [-0.29, 0.35]	0.85 ^{NS}	1.17 [0.83, 1.51]	< .001	0.07 [-0.27, 0.41]	0.68 ^{NS}
Stimulus sensitivity	0.03 [0.03, 0.03]	< .001	0.02 [0.02, 0.02]	< .001	0.02 [0.02, 0.02]	< .001
Stimulus specificity	0.01 [0.01, 0.02]	< .001	0.01 [0.01, 0.01]	< .001	0.02 [0.01, 0.02]	< .001
Stimulus fairness (favours older)	-0.23 [-0.37, -0.1]	< .001	-0.69 [-0.84, -0.55]	< .001	-0.31 [-0.44, -0.18]	< .001
Stimulus fairness (favours younger)	-0.40 [-0.54, -0.27]	< .001	-0.81 [-0.96, -0.66]	< .001	-0.40 [-0.54, -0.27]	< .001
Respondent PGSI score	0.03 [-0.02, 0.08]	0.27 ^{NS}	0.02 [-0.04, 0.07]	0.55 ^{NS}	0.01 [-0.04, 0.07]	0.62 ^{NS}
Respondent age group (18-29)	0.09 [-0.33, 0.5]	0.69 ^{NS}	0.17 [-0.27, 0.61]	0.44 ^{NS}	0.10 [-0.34, 0.54]	0.65 ^{NS}
Respondent age group (50+)	-0.50 [-1.05, 0.04]	0.07 ^{NS}	-0.75 [-1.33, -0.18]	0.01	-0.5 [-1.08, 0.08]	0.10 ^{NS}
Favours older X participants 18-29	-0.22 [-0.45, 0.01]	0.06 ^{NS}	-0.31 [-0.56, -0.06]	0.01	-0.21 [-0.43, 0.02]	0.07 ^{NS}
Favours younger X participants 18-29	-0.13 [-0.36, 0.10]	0.27 ^{NS}	-0.34 [-0.59, -0.10]	0.01	-0.16 [-0.39, 0.06]	0.16 ^{NS}
Favours older X participants 50+	0.06 [-0.24, 0.36]	0.70 ^{NS}	0.07 [-0.26, 0.39]	0.69 ^{NS}	-0.01 [-0.31, 0.30]	0.97 ^{NS}
Favours younger X participants 50+	0.15 [-0.15, 0.45]	0.32 ^{NS}	0.13 [-0.19, 0.46]	0.42 ^{NS}	0.06 [-0.25, 0.36]	0.72 ^{NS}
	$\mathbf{R^2_{Level\ 1}}$	0.26	$\mathbf{R^2_{Level\ 1}}$	0.18	$\mathbf{R^2_{Level\ 1}}$	0.23
	$\mathbf{R^2_{Total}}$	0.49	$\mathbf{R^2_{Total}}$	0.43	$\mathbf{R^2_{Total}}$	0.50

Note: "Stimulus"-related sensitivity, specificity, and fairness refer to manipulations in depicted stimuli (see Figure 1). X indicates interaction term for levels of the participant age group and fairness condition factors. ^{NS} indicates statistical non-significance given a conventional α-level = .05. R^2 = pseudo coefficient of determination given model level. Regression coefficients (*b*) are unstandardized.

Figure 2. Relationships between model performance metrics and performance ratings



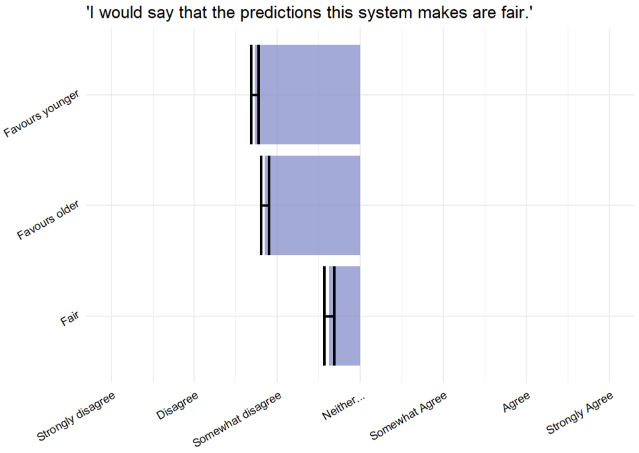
Note: Data points show ratings of the classification performance of proposed AI systems by all participants. Linear trends are fit for both models' depicted sensitivities and specificities. Points have been jittered for visibility

RMSE was 1.28 for this model, indicating that the average model prediction was 1.28 out of 7 rating points away from the true ratings given. A modest group (139 out of 3809) of cases had goodness ratings predicted to be at least one RMSE above "somewhat agree," and together constituted the family of cases whose performance should generally be considered to be "performing well." Indeed, the true ratings of these systems included 96 cases in agreement, 19 neutral, and only 24 in disagreement (of which only one was "strongly disagree"). Twenty-seven different participants were represented among these well-performing cases, varying in both age group ($n_{18\text{-}29} = 10$, $n_{30\text{-}49} = 15$, $n_{50+} = 2$) and self-reported problem gambling status ($n_{PGSI=0} = 4$, $n_{PGSI=1\text{-}2} = 8$, $n_{PGSI=3\text{-}7} = 10$, $n_{PGSI\geq 8} = 5$). A majority (70 out of 139) of these well-performing cases depicted an average FNR of only 10% across age groups; the highest model sensitivity shown. Indeed, the mean sensitivity of all well-performing cases was 80.43%. A plurality of these cases (55 out of 139) depicted an average FPR of only 10% across age groups; the highest specificity shown. The mean specificity of all well-performing cases was 70.36%. The majority (80 out of 139) of well-performing cases depicted fair classification performance across age groups.

Ratings of system fairness

Participants' ratings of whether depicted detection systems appeared to treat people fairly depended on multiple factors. Model fit was strong overall ($R^2_{Total} = 0.43$; Table 2B). The sensitivity, specificity, and fairness manipulations impacted fairness ratings such that highly sensitive ($b = 0.02$, $p < .001$) and highly specific ($b = 0.01$, $p < .001$) systems were rated more fair, and systems which favoured older (*b* = -0.69, *p* < .001) or younger (*b* = -0.81, *p* < .001) people were rated less fair. Thus, participants (rightly) believed that unbalanced classification performance between groups was unfair. Additionally, participants in the 55+ age group rated systems as less fair across the board (*b* = -0.75, *p* = .01), while participants in the 18-29 age group rated systems as less fair when they specifically depicted unfairness favouring older (*b* = -0.31, *p* = .01) or younger (*b* = -0.34, *p* = .01) online gamblers. These effects may explain why even systems depicted as fair were rated as less than fair on average (Figure 3).

Figure 3. Relationships between fairness condition and fairness ratings



Note: Bars represent standard errors. ‘Fair’ systems (Y-axis) are those which displayed no clear advantage for either age group

RMSE was 1.38 for this model, indicating that the average model prediction was 1.38 out of 7 rating points away from the true ratings given. A modest group (126 out of 3836) of cases had system fairness ratings predicted to be at least one RMSE above “somewhat agree,” and together constituted the family of cases whose performance should generally be considered ‘fair.’ Indeed, the true ratings of these systems included 104 cases in agreement, 11 neutral, and 11 in disagreement (of which only two were “strongly disagree”). Thirty different participants were represented among these well-performing cases, varying in both age group ($n_{18\text{-}29}$ = 13, $n_{30\text{-}49}$ = 15, n_{50+} = 2) and self-reported problem gambling status ($n_{PGSI=0}$ = 5, $n_{PGSI=1\text{-}2}$ = 7, $n_{PGSI=3\text{-}7}$ = 15, $n_{PGSI\geq8}$ = 3). A plurality (57 out of 126) of these putatively fair cases depicted an average FNR of only 10% across age groups. The mean sensitivity of all putatively fair cases was 73.89%. A plurality of these cases (59 out of 126) depicted an average FPR of only 10% across age groups. The mean specificity of all putatively fair cases was 68.65%. The majority (94 out of 126) of cases depicted fair classification performance across age groups.

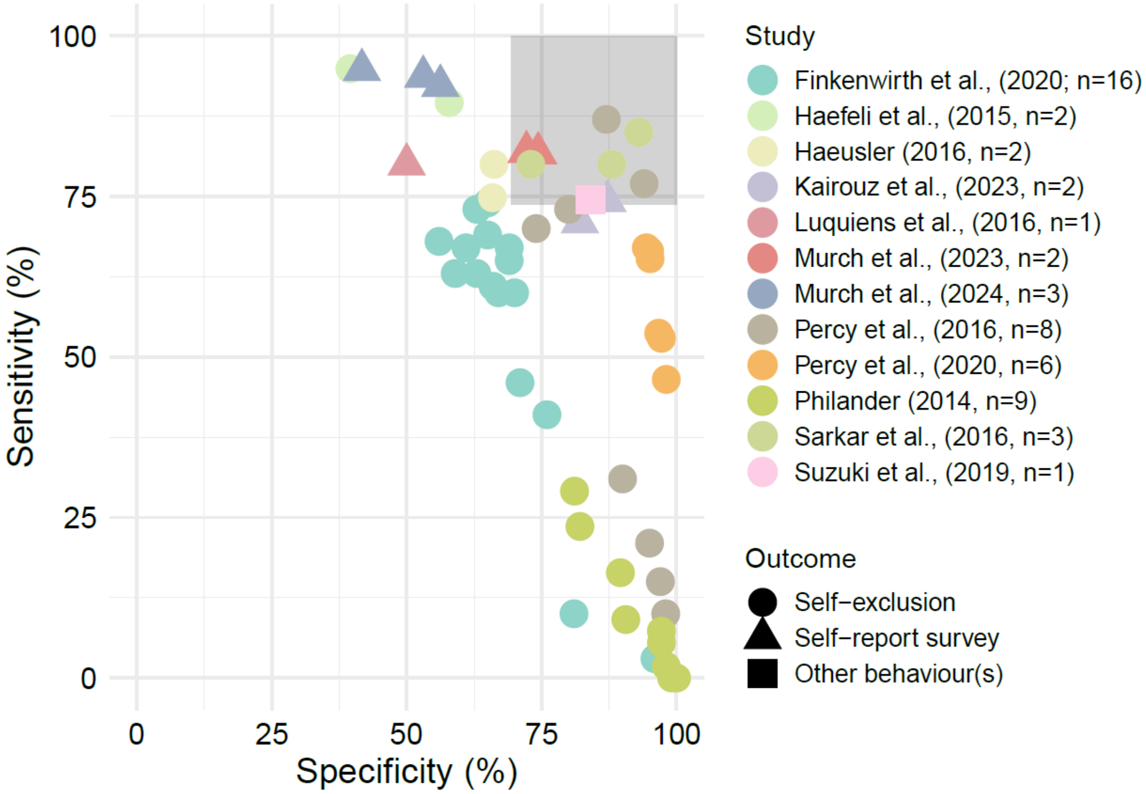
Ratings of real-world readiness

Participants' ratings of whether the depicted systems were ready for use in detecting at-risk gamblers in the real world depended on the experiment's sensitivity, specificity, and fairness conditions. Model fit was strong overall ($R^2_{Total} = 0.50$; Table 2C). The sensitivity and specificity conditions affected participants real-world readiness ratings such that more-sensitive ($b = 0.02, p < .001$) and more-specific ($b = 0.02, p < .001$) systems were rated as more ready for real-world use. Systems were rated significantly less ready for real-world use when they depicted unfairness favouring older ($b = -0.31, p < .001$) or younger ($b = -0.40, p < .001$) people. Readiness ratings did not significantly depend on participants' PGSI scores, their own age, or the interaction between their age group and the depicted fairness condition. Thus, for a system to be considered ready for real-world use, it should meet some standard for all of sensitivity, specificity, and fairness.

RMSE was 1.26 for this model, indicating that the average model prediction was 1.26 out of 7 rating points away from the true ratings given. A modest group (132 out of 3792) of cases had system fairness ratings predicted to be at least one RMSE above "somewhat agree," and together constituted the family of cases whose performance should generally be considered 'ready for real-world use.' Indeed, the true ratings of these systems included 94 cases in agreement, 19 neutral, and 19 in disagreement (of which only one was "strongly disagree"). Twenty two different participants were represented among these well-performing cases, varying in both age group ($n_{18\text{-}29} = 9$, $n_{30\text{-}49} = 11$, $n_{50+} = 2$) and self-reported problem gambling status ($n_{PGSI=0} = 4$, $n_{PGSI=1\text{-}2} = 5$, $n_{PGSI=3\text{-}7} = 9$, $n_{PGSI\geq 8} = 4$). A plurality (55 out of 132) of these real-world ready cases depicted an average FNR of only 10% across age groups. The mean sensitivity of all real-world ready cases was 73.71%. A plurality of these cases (52 out of 132) depicted an average FPR of only 10% across age groups. The mean specificity of all 'real-world ready' cases was 69.39%. The majority (70 out of 132) of real-world ready cases depicted fair classification performance across age groups.

Using these percentages to circumscribe a minimum standard for real world use alongside previously-described models in the gambling area shows that most existing systems did not – at the time of their initial validation – meet the minimum sensitivity and specificity performance standards desired by online gamblers (Figure 4).

Figure 4. Judgments of AI systems' readiness for real-world use



Note: Points depict the reported performance levels of gambling-relevant AI systems previously reported by multiple authors. Shaded region shows means from the sample of visualization task data where participants' ratings of real-world readiness were predicted to be "somewhat agree" plus one Root Mean Squared Error (RMSE). Thus, the shaded region represents performance levels that on average were at least somewhat agreeable to Canadian online gamblers. The points falling within the shaded region – if correctly fit, correctly reported, and fair – could be considered at least somewhat ready for real-world use. Colors distinguish studies while point shapes distinguish classified outcomes. *n* = number of models displayed on the graph. Data used in this plot are archived (**Note 1**)

Discussion

We conducted a data visualization experiment using an adequately powered sample of Canadian online gamblers in order to understand more about this population's views on automated systems that seek to detect and prevent harms related to online gambling. Such systems (many of them using AI technologies) may become widespread in gambling jurisdictions as evidence accumulates showing their potential to distinguish majorities of at-risk and lower-risk gamblers. The participant sample was comprised of a high number of at-risk gamblers compared to the Canadian national average (Williams et al., 2021) and rates seen among users of Canadian online gambling sites (Murch et al., 2023). This is likely due to our having screened for online gamblers ahead of their participation, but provided a strong analytical basis for understanding the perspectives of both higher- and lower-risk online gamblers. Participants in the experiment were asked to judge the performance, fairness, and real-world readiness of a series of hypothetical gambling harm detection systems that varied in terms of their displayed False Negative Rate (FNR; 1-sensitivity), False Positive Rate (FPR; 1-specificity), and fairness with respect to different age groups. We hypothesized that participants' ratings would vary along these dimensions, that ratings of model performance would be differently affected by FNR and FPR, that models depicting unfairness would be rated as less fair, and that participants' own age and experience of gambling problems would impact these ratings. Hypotheses 1, 1.1, and 1.2 were supported, while hypothesis 1.3 received partial support.

Participants' responses to the question of hypothetical systems' performance depended both on the overall classification performance depicted in terms of FNR and FPR, and on the depicted fairness of classifications being made. Two facts are crucial to notice with this result: first, participants were more concerned with models' depicted sensitivity than their specificity, rewarding very sensitive models and punishing very insensitive models (Figure 2). Thus, although correct classification appeared to matter for both those experiencing and those not experiencing gambling problems, participants were most concerned with sensitively detecting problem gamblers. Second, models that were depicted as unfair were rated as worse-performing even when their overall classification performance matched that of systems depicted as fair. Thus, participants' ratings of classification performance are contingent upon whether the system appeared fair or not. From this model, we computed and proposed minimally-acceptable standards for well-performing systems that indicated a need for: (1) fairness across ages, (2) sensitivity ≈ 80%, and (3) specificity ≈ 70%.

With regards to the question of models' fairness, we found similar trends for sensitivity, specificity, and depicted fairness. Once again, participants' ratings favoured systems with high sensitivity, high specificity, and fairness across age groups which suggested a preference for sensitively detecting problem gambling. This model additionally showed age group effects for those aged 55+, who appeared to rate all systems as less fair than other age groups. Similar trends were seen for those in the 18 – 29 age group, but only when the depicted detection system actually showed age-related discrepancies. Together, these results suggest that "fair" systems ought to be fair across age groups, and able to correctly classify about fifteen in twenty people with gambling problems as well as about seven in ten people without gambling problems. They also suggest that concepts of fairness in automated detection may vary between generations. Further research is needed to understand the nature of trait differences found here. From this model, we nevertheless computed and proposed minimally-acceptable standards for 'fair' systems, indicating a need for: (1) fairness across ages, (2) sensitivity ≈ 74%, and (3) specificity ≈ 69%.

With regards to the question of models' readiness for real-world deployment, we again found that participants' ratings were positively related to sensitivity and specificity, and negatively related to apparent unfairness across ages. It is thus important to Canadian online gamblers that systems which seek to detect online gambling problems be both fair across age groups, and strongly performant for both those experiencing past-year gambling problems and those not experiencing past-year gambling problems. Based on these data, we thus propose minimally-acceptable standards for real-world deployment that demonstrate: (1) fairness across ages, (2) sensitivity ≈ 74%, and (3) specificity ≈ 69%. Looking across a variety of existing AI-based gambling harm detection systems (Figure 4) described in the literature and recent systematic analyses (Ghaharian et al., 2022), it is clear that few existing systems have reached these standards for sensitivity and specificity, and almost none have assessed classification fairness across age groups or other sociodemographic dimensions (but see Percy et al., 2020; Murch et al., 2024a).

Since relatively few systems appear to meet the performance standards indicated by participants in this study, these findings thus call into question the readiness of AI technology for the purposes of detecting and preventing online gambling-related harms. Although some AI systems may perform well at the time of their initial validation, this performance may diminish over time (Murch et al., 2024b), and we have yet to see any system perform acceptably well over time while also demonstrating classification fairness with respect to age. This is not to say that AI technologies cannot ever be used for the detection of at-risk online gamblers; just that they have not yet shown they can perform in a manner that would be deemed even somewhat agreeable to online gamblers, the group who stands to be most affected by their (mis)behaviour. This is alarming from both a legal and an ethical standpoint, as *fair performance consistent with end users values* is a cornerstone in both ethical frameworks for responsible AI development and proposed AI legislation (Canada Bill C-27, 2023; Montréal Institute for Learning Algorithms, 2018).

That said, the system performance levels outlined here should be thought of as minimum thresholds and not end goals. Although a future system's classification errors may be infrequent enough to achieve acceptability for real world use, the harms caused by any errors that did occur could be serious. Any lack of sensitivity implies that some number of truly-at-risk people would be passed over by a given detection system. These individuals would not receive any system-directed prevention materials such as informational support or referrals to treatment services, prolonging their gambling harms, reducing the overall efficacy of the prevention strategy, and creating potential health disparities between demographic groups (Murch et al., 2024a). On the other hand, any lack of specificity implies that some number of truly-lower-risk people would be incorrectly flagged as at-risk. As a result, limited prevention resources (e.g., referrals to speak with a counsellor) may be applied inefficiently to lower-risk individuals and may run out before all higher-risk cases are seen. Separately, gambling platforms may be hesitant to deploy a system that curtails their operations erroneously; supposing that a detection system stops promotional advertising to all users predicted to be at-risk, false positives could be seen by the operator as unduly hampering their commercial efforts. More broadly, any errors (acceptable or not) may cut against the perceived credibility of a system. If these doubts were spread on forums or social media, even the system's correct predictions could be undermined in the minds of end users. So, although future AI-based harm detection systems may be acceptably performant to the standards outlined here, developers of these systems must continue working towards maximal classification performance to reduce the numerous serious externalities that may result from classification errors.

Several limitations to these results should be noted. First, a convenience sampling method was employed, raising the possibility that users of the online survey platform are not a representative subset of all online gamblers in Canada. In particular, the sample did not find a broad representation of French-speaking or gender-diverse Canadians, and these facts prevent fully generalizing our findings to the wider population of Canadian online gamblers. Second, due to constraints on participants' time and the total sample size afforded by research funds, we could not consider multiple aspects of sociodemographic fairness beyond participants' age group. Third, the stimuli in this study did not anticorrelate depicted sensitivity and specificity values in a manner consistent with many AI-based classification models in the field. Reflecting this anticorrelation could impact preferences for a given model's performance, fairness, or real-world readiness. We additionally could not consider multiple degrees of (un)fairness while maintaining only 36 counterbalanced stimuli per participant, and thus we fixed the degree of unfairness depicted in the experiment stimuli at a uniform 20% (10% advantage to one age group + 10% disadvantage to the other) discrepancy between age groups. These results thus cannot identify acceptable levels of unfairness across age groups or other sociodemographic factors.

An additional limitation of this work concerns our omission of comprehension checks. Although the obtained results suggest that participants generally responded rationally based on the stimuli they were shown (stimuli depicting poor classification performance were rated as less performant than those with strong classification performance, stimuli depicting unfairness were rated less fair, etc.), it remains possible that all participants did not fully grasp or differentiate the complex concepts being presented. This is one potential explanation for the high collinearity between this study's three outcome variables both within participants and within task stimuli. Finally, the hypotheses tested here were exploratory and not pre-registered. Replication of these findings is warranted.

To our knowledge, this is the first study to use data visualization as a means of querying online gamblers views on automated harm detection systems. In terms of future research, several interesting avenues exist. Straightforwardly, it is likely that the exact manner of information presentation impacts the views elicited or the certainty of their expression. By presenting model performance information in numeric terms rather than as percentages, reported views may shift in a manner consistent with the lower prevalence of at-risk compared to lower-risk gamblers typically observed. Similarly, additional insight may be revealed by providing and explaining alternate performance metrics (F1 score, overall accuracy, etc.). Another interesting avenue of research would be to systematically manipulate task instructions to modulate participants understanding of what “gambling problems”, “counselors”, or “help” entailed. Yet another interesting follow-up could examine the role of transparency in end users’ perceptions; do online gamblers prefer a better-performing detection system that relies on ‘black box’ modelling algorithms, or would they trade some amount of classification performance for the chance to understand exactly how any individual decision was reached? The limited nature of this first study did not allow consideration of these possibilities, but it is likely that considerations related to: (1) whom a given detection system applies, and (2) what will be done with it, impact online gamblers’ views on whether it is ready for real-world use.

Conclusion

Much interest and hype has surrounded the use of Artificial Intelligence for detecting and preventing online gambling-related harms. Arguably the most important question regarding the utility of AI-based systems for online gambling harm detection is: “does this technology work well enough to be used on those who stand to be benefitted (or harmed) by it?” Here, we showed that Canadian online gamblers have sensible and realistic desires for the performance of these automated detection systems which may impact their lives (Marionneau et al., 2025). From these views, we identified minimally-acceptable thresholds for current and existing models’ sensitivities, specificities, and fairness with respect to older and younger persons. These thresholds are separated along the relevant questions of whether or not a system performs well, seems fair, and appears ready for real-world use.

In a broader sense, these results provide a roadmap for developing AI-based harm reduction materials in a manner that centers end users’ experiences and preferences. This is relevant to recognized behavioural addictions including video games and gambling platforms which collect sufficient data that is both individualized and labelled in a way that enables the application of classification algorithms. It is also clearly relevant in subclinical or prevention-oriented cases where at-risk individuals may be identified and reached with early intervention materials. Separately, such systems may be beneficial for the purposes of product safety or regulator oversight, estimating the relative harmfulness of specific online gambling activities or the number of at-risk people who should have been reached for duty of care interventions within some period. Finally, this approach may be useful if employed alongside large datasets from financial service providers for the prevention of substance use disorders wherever legalized substances (e.g., alcohol, tobacco, and cannabis) could be identified through transactions at certain businesses over time. In each of these cases, we propose that the most responsible path towards the development of AI-based harm prevention systems is to: (1) systematically fit and validate a potential detection model (e.g., Murch, Kairouz, Dauphinais, et al., 2023), (2) routinely re-validate the model to determine whether its performance is drifting over time (Lu et al., 2019; Murch et al., 2024b), (3) identify and remedy any sociodemographic biases exhibited by the model in a 'fair' manner (Murch et al., 2024a), and (4) conduct independent testing with end users (as above) to identify minimally-acceptable standards of performance. Ideally, in the coming years these methods for automated detection – AI-based or otherwise – will evolve in tandem with improved materials and procedures for personalized intervention with individuals predicted to experience varying degrees of online gambling problems.

Note 1: https://osf.io/n9sf4/

Supplementary Materials for "Online Gamblers' Preferences for Performance and Fairness in Artificial Intelligence Systems for Gambling Harm Detection"

A checklist for conjoint analysis applications in health care (Bridges et al., 2011)

1. Was a well-defined research question stated and is conjoint analysis an appropriate method for answering it?
    - ✓ 1.1 Were a well-defined research question and a testable hypothesis articulated?
    - ✓ 1.2 Was the study perspective described, and was the study placed in a particular decision-making or policy context?
    - ✓ 1.3 What is the rationale for using conjoint analysis to answer the research question?
        - o *Note:* The study seeks to determine what levels of classification error and unfairness could be considered permissible in the use of automated gambling harm detection systems. This is explained in more detail in the manuscript.
2. Was the choice of attributes and levels supported by evidence?
    - 2.1 Was attribute identification supported by evidence (literature reviews, focus groups, or other scientific methods)?

- *Note:* No prior study has examined end user perceptions of automated gambling harm detection tools. Attributes were selected on the grounds that they are three of the most theoretically relevant aspects of the functioning of any automated detection tool (sensitivity, specificity, and fairness). They are not a comprehensive set, and this is a noted limitation.

- ✓ 2.2 Was attribute selection justified and consistent with theory?
    - *Note:* See 2.1 above.
- ✓ 2.3 Was level selection for each attribute justified by the evidence and consistent with the study perspective and hypothesis?
    - *Note:* Attributes ranging from 0-100% were included. To keep the total number of stimuli within reason, specific values between 0 and 100% were selected so that results could interpolate well across the full percentage scale.

3. Was the construction of tasks appropriate?
    - ✓ 3.1 Was the number of attributes in each conjoint task justified (that is, full or partial profile)?
        - *Note:* The statistical modelling strategy would not have worked well without a full profile design.
    - ✓ 3.2 Was the number of profiles in each conjoint task justified?
        - *Note:* Not applicable – a full-profile design was used.
    - ✓ 3.3 Was (should) an opt-out or a status-quo alternative (be) included?
        - *Note:* A status-quo alternative was not included. Rather, the task framed automated detection tools as not-currently in use, and asked participants to evaluate whether they should be put into use.
4. Was the choice of experimental design justified and evaluated?
    - ✓ 4.1 Was the choice of experimental design justified? Were alternative experimental designs considered?
    - ✓ 4.2 Were the properties of the experimental design evaluated?
        - *Note:* Attributes were selected and counterbalanced in a manner that
            - Minimized their intercorrelation.
            - Maintained balance in the number of examples for each level of each independent variable.
            - Restricted impossible circumstances (e.g., percentage values < 0%).
    - ✓ 4.3 Was the number of conjoint tasks included in the data-collection instrument appropriate?
        - *Note:* One task only.
5. Were preferences elicited appropriately, given the research question?
    - ✓ 5.1 Was there sufficient motivation and explanation of conjoint tasks?

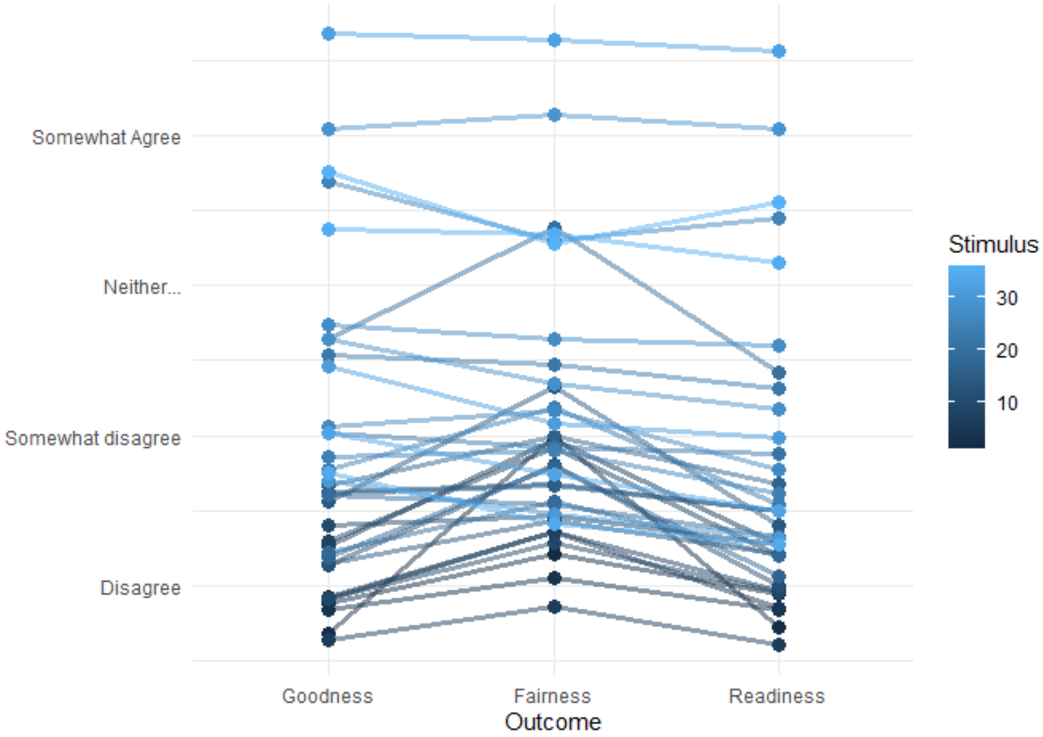
        - *Note:* Participants were provided with a full explanation, and were given hints that helped them understand how to read data visualizations.
    - ✓ 5.2 Was an appropriate elicitation format (that is, rating, ranking, or choice) used? Did (should) the elicitation format allow for indifference?
        - *Note:* Typical seven-point Likert scales were used with midpoint values present for indifferent responding.
    - ✓ 5.3 In addition to preference elicitation, did the conjoint tasks include other qualifying questions (for example, strength of preference, confidence in response, and other methods)?
        - *Note:* Strength of preference was built into the Likert scale ratings employed.

6. Was the data collection instrument designed appropriately?
    - ✓ 6.1 Was appropriate respondent information collected (such as sociodemographic, attitudinal, health history or status, and treatment experience)?
        - *Note:* Yes. See Table 1.
    - ✓ 6.2 Were the attributes and levels defined, and was any contextual information provided?
        - *Note:* Attributed were defined in practical terms with their visual corollaries pointed out in the task instructions.
    - ✓ 6.3 Was the level of burden of the data-collection instrument appropriate? Were respondents encouraged and motivated?
        - *Note:* Participants were compensated competitively for the recruitment platform selected. Inattentive participants were excluded from analysis.
7. Was the data-collection plan appropriate?
    - ✓ 7.1 Was the sampling strategy justified (for example, sample size, stratification, and recruitment)?
        - *Note:* Sample size was determined *a priori* to achieve adequate power.
    - ✓ 7.2 Was the mode of administration justified and appropriate (for example, face-to-face, pen-and-paper, web-based)?
        - *Note:* Online surveying is standard in the field.
    - ✓ 7.3 Were ethical considerations addressed (for example, recruitment, information and/or consent, compensation)?
8. Were statistical analyses and model estimations appropriate?
    - ✓ 8.1 Were respondent characteristics examined and tested?
        - *Note:* Reported models included relevant traits.
    - ✓ 8.2 Was the quality of the responses examined (for example, rationality, validity, reliability)?
        - *Note:* Results were consistent with rational responding. Stimuli depicting better-performing systems were rated as better-performing. Stimuli depicting unfairness were rated as less fair.

- ✓ 8.3 Was model estimation conducted appropriately? Were issues of clustering and subgroups handled appropriately?
    - o *Note:* Models were fit using standard procedures, and model assumptions were checked.

9. Were the results and conclusions valid?
    - ✓ 9.1 Did study results reflect testable hypotheses and account for statistical uncertainty?
    - ✓ 9.2 Were study conclusions supported by the evidence and compared with existing findings in the literature?
        - o *Note:* No existing findings for this topic are present in the gambling field.
    - ✓ 9.3 Were study limitations and generalizability adequately discussed?
10. Was the study presentation clear, concise, and complete?
    - ✓ 10.1 Was study importance and research context adequately motivated?
    - ✓ 10.2 Were the study data-collection instrument and methods described?
    - ✓ 10.3 Were the study implications clearly stated and understandable to a wide audience?

Figure S.1. Mean outcome variable ratings by stimulus



Note: Mean ratings for the study's three dependent variables, each averaged across participants for within the experiment's 36 stimuli.

Funding

No funding was received for the study. The authors wish to acknowledge the support of Lives Lived Well and the clients whose de-identified data made this research possible.

Statement of Competing Interests

None.

Author's contributions

All authors contributed to the study conception, design, and literature review. Data collection and data analysis were performed by N.K. and H.S. The first draft of the manuscript was written by N.K., H.S., and R.W., with G.D. and S.Z. contributing to revisions and improvements. All authors read and approved the final manuscript.

Ethics Approval

Human research ethics exemption was granted by the University of Queensland Human Research Ethics Committee (2023/HE001205) on 17/07/2023 for the project titled 'Alcohol and other drug moderators of the relationship between negative emotional states, emotional impulsivity, and problematic gambling.

Research Promotion

This study explored how emotional impulsivity contributes to gambling harm and how this relationship is intensified by methamphetamine use. Findings highlight that both negative and positive emotional impulsivity predict gambling problems, with methamphetamine uniquely moderating this link. The results have implications for integrated treatment approaches in AOD and gambling services.

References

Akhter, S. A. (2017). *Using machine learning to predict potential online gambling addicts* . [Aalto University]. https://aaltodoc.aalto.fi/server/api/core/bitstreams/39e46281-8631-46f6-8381-8771f9f5dfe9/content

Auer, M., & Griffiths, M. D. (2022). Using artificial intelligence algorithms to predict self-reported problem gambling with account-based player data in an online casino setting. *Journal of Gambling Studies*. https://doi.org/10.1007/s10899-022-10139-1

Badillo, S., Banfai, B., Birzele, F., Davydov, I. I., Hutchinson, L., Kam-Thong, T., Siebourg-Polster, J., Steiert, B., & Zhang, J. D. (2020). An Introduction to Machine Learning. *Clinical Pharmacology & Therapeutics*, *107*(4), 871–885. https://doi.org/10.1002/cpt.1796

Bartoń, K. (2024). *MuMIn: Multi-Model Inference*. https://CRAN.R-project.org/package=MuMIn

Benjamin, R. (2019). *Race After Technology*. Polity.

Braverman, J., & Shaffer, H. J. (2012). How do gamblers start gambling: Identifying behavioural markers for high-risk internet gambling. *European Journal of Public Health*, *22*(2), 273–278. https://doi.org/10.1093/eurpub/ckp232

Bridges, J. F. P., Hauber, A. B., Marshall, D., Lloyd, A., Prosser, L. A., Regier, D. A., Johnson, F. R., & Mauskopf, J. (2011). Conjoint Analysis Applications in Health—a Checklist: A Report of the ISPOR Good Research Practices for Conjoint Analysis Task Force. *Value in Health*, *14*(4), 403–413. https://doi.org/10.1016/j.jval.2010.11.013

Canada Bill C-27, C–27, House of Commons 44th Parliament, 1st Session (2023). https://www.parl.ca/legisinfo/en/bill/44-1/c-27

Corbett-Davies, S., & Goel, S. (2018). *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning* (No. arXiv:1808.00023). arXiv. https://doi.org/10.48550/arXiv.1808.00023

Currie, S. R., Hodgins, D. C., & Casey, D. M. (2013). Validity of the Problem Gambling Severity Index Interpretive Categories. *Journal of Gambling Studies*, *29*(2), 311–327. https://doi.org/10.1007/s10899-012-9300-6

Das, M. (2021). Fantasy sports and gambling regulation in the Asia-Pacific. *The International Sports Law Journal*, *21*(3), 166–179. https://doi.org/10.1007/s40318-021-00198-8

Dellis, A., Sharp, C., Hofmeyr, A., Schwardmann, P. M., Spurrett, D., Rousseau, J., & Ross, D. (2014). Criterion-related and construct validity of the Problem Gambling Severity Index in a sample of South African gamblers. *South African Journal of Psychology*, *44*(2), 243–257. https://doi.org/10.1177/0081246314522367

Ferris, J., & Wynne, H. (2001). The Canadian Problem Gambling Index: Final report. In *Canadian Centre on Substance Abuse* (p. 38). https://doi.org/10.1007/s10899-010-9224-y
Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R* (1st edition). SAGE Publications Ltd.

Finkenwirth, S., MacDonald, K., Deng, X., Lesch, T., & Clark, L. (2020). Using machine learning to predict self-exclusion status in online gamblers on the PlayNow.com platform in British Columbia. *International Gambling Studies*, *21*(2), 220–237. https://doi.org/10.1080/14459795.2020.1832132

Gainsbury, S. M. (2015). Online Gambling Addiction: The Relationship Between Internet Gambling and Disordered Gambling. *Current Addiction Reports*, *2*(2), 185–193. https://doi.org/10.1007/s40429-015-0057-8

Ghaharian, K., Abarbanel, B., Phung, D., Puranik, P., Kraus, S., Feldman, A., & Bernhard, B. (2022). Applications of data science for responsible gambling: A scoping review. *International Gambling Studies*, *0*(0), 1–24. https://doi.org/10.1080/14459795.2022.2135753

Gooding, N. B., & Williams, R. J. (2023). Are There Riskier Types of Gambling? *Journal of Gambling Studies*. https://doi.org/10.1007/s10899-023-10231-0

Haefeli, J., Lischer, S., & Haeusler, J. (2015). Communications-based early detection of gambling-related problems in online gambling. *International Gambling Studies*, *15*(1), 23–38. https://doi.org/10.1080/14459795.2014.980297

Haeusler, J. (2016). Follow the money: Using payment behaviour as predictor for future self-exclusion. *International Gambling Studies*, *16*(2), 246–262. https://doi.org/10.1080/14459795.2016.1158306

Holtgraves, T. (2009). Evaluating the problem gambling severity index. *Journal of Gambling Studies*, *25*(1), 105–120. https://doi.org/10.1007/s10899-008-9107-7

Kairouz, S., Costes, J. M., Murch, W. S., Doray-Demers, P., Carrier, C., & Eroukmanoff, V. (2023). Enabling New Strategies to Prevent Problematic Online Gambling: A machine learning approach for identifying at-risk online gamblers in France. *International Gambling Studies*. https://doi.org/10.1080/14459795.2022.2164042

Ladouceur, R., Shaffer, P., Blaszczynski, A., & Shaffer, H. J. (2017). Responsible gambling: A synthesis of the empirical evidence. *Addiction Research and Theory*, *25*(3), 225–235. https://doi.org/10.1080/16066359.2016.1245294

LaPlante, D. A., Nelson, S. E., & Gray, H. M. (2014). Breadth and depth involvement: Understanding Internet gambling involvement and its relationship to gambling problems. *Psychology of Addictive Behaviors*, *28*(2), 396–403. https://doi.org/10.1037/a0033810

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2019). Learning under Concept Drift: A Review. *IEEE Transactions on Knowledge and Data Engineering*, *31*(12), 2346–2363. IEEE Transactions on Knowledge and Data Engineering. https://doi.org/10.1109/TKDE.2018.2876857

Luquiens, A., Tanguy, M.-L., Benyamina, A., Lagadec, M., Aubin, H. J., & Reynaud, M. (2016). Tracking online poker problem gamblers with player account-based gambling data only. *International Journal of Methods in Psychiatric Research*, *25*(4), 332–342. https://doi.org/10.1002/mpr.1510

Marionneau, V., Ristolainen, K., & Roukka, T. (2025). Duty of care, data science, and gambling harm: A scoping review of risk assessment models. *Computers in Human Behavior Reports*, 100644. https://doi.org/10.1016/j.chbr.2025.100644

Miller, N. V., Currie, S. R., Hodgins, D. C., & Casey, D. (2013). Validation of the problem gambling severity index using confirmatory factor analysis and rasch modelling. *International Journal of Methods in Psychiatric Research*, *22*(3), 245–255. https://doi.org/10.1002/mpr.1392

Montréal Institute for Learning Algorithms. (2018). *Montréal Declaration for a Responsible Development of Artificial Intelligence* (pp. 1–21). https://www.montrealdeclaration-responsibleai.com/

Murch, W. S., Kairouz, S., Dauphinais, S., Picard, E., Costes, J.-M., & French, M. (2023). Using machine learning to retrospectively predict self-reported gambling problems in Quebec. *Addiction*. https://doi.org/10.1111/add.16179

Murch, W. S., Kairouz, S., & French, M. (2024a). Comparing 'fair' machine learning models for detecting at-risk online gamblers. *International Gambling Studies*, *0*(0), 1–23. https://doi.org/10.1080/14459795.2024.2412051

Murch, W. S., Kairouz, S., & French, M. (2024b). Establishing the Temporal Stability of Machine Learning Models That Detect Online Gambling-Related Harms. *Computers in Human Behavior Reports*, 100427. https://doi.org/10.1016/j.chbr.2024.100427

Neuwirth, E. (2014). *RColorBrewer: ColorBrewer Palettes* [Computer software]. https://cran.r-project.org/package=RColorBrewer

Olason, D. T., Kristjansdottir, E., Einarsdottir, H., Haraldsson, H., Bjarnason, G., & Derevensky, J. L. (2011). Internet Gambling and Problem Gambling Among 13 to 18 Year Old Adolescents in Iceland. *International Journal of Mental Health and Addiction*, *9*(3), 257–263. https://doi.org/10.1007/s11469-010-9280-7

Percy, C. (2020). Lessons Learned from Problem Gambling Classification: Indirect Discrimination and Algorithmic Fairness. *AI4SG@ AAAI Fall Symposium*. https://ceur-ws.org/Vol-2884/paper_107.pdf

Percy, C., França, M., Dragičević, S., & d'Avila Garcez, A. (2016). Predicting online gambling self-exclusion: An analysis of the performance of supervised machine learning models. *International Gambling Studies*, *16*(2), 193–210. https://doi.org/10.1080/14459795.2016.1151913

Petry, N. M. (2006). Internet gambling: An emerging concern in family practice medicine? *Family Practice*, *23*(4), 421–426. https://doi.org/10.1093/fampra/cml005

Philander, K. S. (2014). Identifying high-risk online gamblers: A comparison of data mining procedures. *International Gambling Studies*, *14*(1), 53–63. https://doi.org/10.1080/14459795.2013.841721

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2019). *nlme: Linear and Nonlinear Mixed Effects Models* [Computer software]. https://cran.r-project.org/package=nlme

Potenza, M. N., Balodis, I. M., Derevensky, J., Grant, J. E., Petry, N. M., Verdejo-garcia, A., & Yip, S. W. (2019). Gambling Disorder. *Nature Reviews Disease Primers*, *5*(51), 1–21. https://doi.org/10.1038/s41572-019-0099-7

R Core Team. (2023). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. http://www.r-project.org/

Reynolds, J., Kairouz, S., Ilacqua, S., & French, M. (2020). Responsible Gambling: A Scoping Review. *Critical Gambling Studies*, *1*(1), 23–39. https://doi.org/10.29173/cgs42

Samuelsson, E., Wennberg, P., & Sundqvist, K. (2019). Gamblers' (mis-)interpretations of Problem Gambling Severity Index items: Ambiguities in qualitative accounts from the Swedish Longitudinal Gambling Study. *NAD Nordic Studies on Alcohol and Drugs*, *36*(2), 140–160. https://doi.org/10.1177/1455072519829407

Shead, W. N., Hodgins, D. C., & Scharf, D. (2008). Differences between Poker Players and Non-Poker-Playing Gamblers. *International Gambling Studies*, *8*(2), 167–178. https://doi.org/10.1080/14459790802139991

Stark, S., & Robinson, J. (2021). Online gambling in unprecedented times: Risks and safer gambling strategies during the COVID-19 pandemic. *Journal of Gambling Issues*, *47*(17), 409–423. https://doi.org/10.4309/jgi.2021.47.17

Statistics Canada. (2019). *Canadian Community Health Survey—Annual Component (CCHS)*. https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3226

Stewart, M., Brown, J. B., Donner, A., McWhinney, I. R., Oates, J., Weston, W. W., & Jordan, J.

(2000). The impact of patient-centered care on outcomes. *The Journal of Family Practice*, *49*(9), 796–804.

Sweeney, L. (2013). *Discrimination in Online Ad Delivery* (SSRN Scholarly Paper No. 2208240). https://doi.org/10.2139/ssrn.2208240

van Berkel, N., Goncalves, J., Russo, D., Hosio, S., & Skov, M. B. (2021). Effect of Information Presentation on Fairness Perceptions of Machine Learning Predictors. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3411764.3445365

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis* [Computer software]. Springer-Verlag. https://ggplot2.tidyverse.org

Williams, R. J., Hann, R. G., Schopflocher, D. P., West, B. L., McLaughlin, P., White, N., King, K., & Flexhaug, T. (2015). *Quinte longitudinal study of gambling and problem gambling*. Ontario Problem Gambling Research Centre. http://hdl.handle.net/10133/3641

Williams, R. J., Leonard, C. A., Belanger, Y. D., Christensen, D. R., el-Guebaly, N., Hodgins, D. C., McGrath, D. S., Nicoll, F., & Stevens, R. M. G. (2021). Gambling and Problem Gambling in Canada in 2018: Prevalence and Changes Since 2002. *Canadian Journal of Psychiatry*, *66*(5), 485–494. https://doi.org/10.1177/0706743720980080

Williams, R. J., & Volberg, R. A. (2014). The classification accuracy of four problem gambling assessment instruments in population research. *International Gambling Studies*, *14*(1), 15–28. https://doi.org/10.1080/14459795.2013.839731

Xuan, Z., & Shaffer, H. (2009). How do gamblers end gambling: Longitudinal analysis of internet gambling behaviors prior to account closure due to gambling related problems. *Journal of Gambling Studies*, *25*(2), 239–252. https://doi.org/10.1007/s10899-009-9118-z

Article Submission: https://jgi.manuscriptmanager.net/