UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
What is addiction? Substance-specific biases in human beliefs and LLMs

Permalink
https://escholarship.org/uc/item/4bt6f9ns

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Authors
Martin Lopez, Maria
Joyner, Keanan
Thompson, Bill

Publication Date
2025

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

What is addiction? Substance-specific biases in human beliefs and LLMs

Maria Martin Lopez, Keanan J. Joyer, and Bill D. Thompson
{mmartinlopez, kjoyner, wdt}@berkeley.edu
Department of Psychology, 2121 Berkeley Way
Berkeley, CA 94701 USA

Abstract

Understanding how individuals conceptualize addiction is an important approach to the study of substance use etiology. We asked participants in a large free-response study of of intuitive conceptualizations of addiction among alcohol and cannabis users and co-users to share their beliefs about the benefits and harms of alcohol and cannabis, and to explain in simple terms what it means to be addicted. Using a frontier language model (ChatGPT-4o) we extracted structured representations of people's beliefs and explanations, assessing the extent to which responses represented 11 clinically relevant diagnostic symptoms from the DSM-5 section on Substance Use Disorders. People's beliefs showed clear substance-specific biases, attributing more clinically relevant symptoms to alcohol than cannabis. A prompt-context manipulation that contextualized participants' substance-neutral explanations as relevant to either cannabis or alcohol revealed evidence sometimes for similar, and for other times opposite direction, substance-specific biases imposed by the ChatGPT annotation process itself.

Keywords: conceptual structure, language, psychopathology

Introduction

Misuse of drugs and alcohol imposes an important public and personal health burden and has become an increasingly prevalent concern in recent years (Degenhardt et al., 2018; Lu, Lopez-Castro, & Vu, 2023). Given the impairing nature of addiction, research on this disorder is a global priority. Understanding how people conceptualize addiction and what it means to be addicted is an important approach to studying addiction's etiology. Our clinical diagnostic term for addiction is a Substance Use Disorder (SUD), which is diagnosed via oral interviews using the diagnostic symptoms defined by the DSM-5 (Douaihy & Daley, 2013) and therefore a person's impairment on a series of defined symptoms are mediated by the language they use to describe their experiences, thoughts, and behaviors surrounding substances (Clark, 2011). While there is a large amount of research on the nature of SUDs there remain key gaps in our understanding of the intuitive conceptualizations of the public about addiction.

There are widely-acknowledged differences in the ways that people perceive addiction to alcohol versus cannabis. While there are very different psychopharmacological effects of each drug there are also common beliefs that are disconnected from the these differences. Research shows that perceptions of harm related to cannabis and alcohol are perceived differently across demographic groups and alcohol is often perceived as more harmful than cannabis (Allen et al., 2018; Leung, Chan, Hides, & Hall, 2020). Such substance-specific biases in perceptions of harm, risk, and morality can directly influence attitudes toward addiction and inform clinical conceptualizations of risk. Our understanding of substance-specific biases in people's intuitive theories of addiction remains limited. Public surveys often rely primarily on quantitative Likert-style ratings, restricting the potential for insight into the underlying cognitive structure of people's beliefs surrounding addiction. One way to gain further insight is to elicit reflection through free-response paradigms. Free-response data can surface qualitative aspects of conceptual structure that are most salient to people, but have traditionally been difficult to analyze in a scalable way.

Innovations surrounding Large Language Models (LLMs) offer the potential to combine the merits of naturalistic free-response user reports with the power of quantitative analysis. However, the use of these models has also raised significant concerns about the potential for model-driven bias (Ranjan, Gupta, & Singh, 2024). LLMs may perpetuate or amplify existing human biases. Studies in related areas have found evidence for amplification of gender, racial, and contextual biases, for example (Ahn & Oh, 2021; Bordia & Bowman, 2019; Hu et al., 2024). LLMs may also impose biases that differ significantly from human understanding, as a consequence of lacking relevant social or ethical context, for example (Morales, Clarisó, & Cabot, 2024; Musolesi, 2024; O'neil, 2017). This potential for bias amplification and misalignment is particularly important for clinically-related topics.

In this study, we explored the structure of people's intuitive understanding of addiction, a relatively abstract concept with concrete clinical relevance. We asked 200 participants who regularly consumed alcohol (and 60% regularly consumed cannabis) to explain the concept of addiction in their own words, and to describe the benefits and harms of alcohol and cannabis as they see them. We used a frontier language model (ChatGPT-4o) to extract structured features from people's free-responses using each of the DSM-5 diagnostic criteria for alcohol and cannabis. For example, we tasked the model with assessing the extent to which any given response included craving as part of the explanation, or tolerance, etc. We explored evidence for systematic and substance-specific biases in the features people construe as most relevant to addiction. We examined whether these biases are reproduced or distorted through the use of a LLM to analyze people's

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

beliefs and explanations, and found that ChatGPT-4o shows bias wherein some symptoms of SUDs were more likely to be represented in responses when the target substance was alcohol compared to cannabis, and for other symptoms, the opposite was true. Overall human biases were higher for both alcohol and cannabis, yet the biases created by the LLM were often misaligned with these.

Methods

Participants

$N = 200$ participants were recruited via Prolific. Twenty-nine were not included in the final sample because their responses failed basic data quality criteria. Our final sample contained N = 171 (US, fluent English) participants between 21 to 30 years of age, and met the minimum approval rate requirements on Prolific. Our sample had a fairly even distribution of people who identified as male (48%) and female (52%). The sample is predominantly White (45. 6%) with significant representation of racial minorities: 26.3% Black, 6.4% Asian, 8.77% Hispanic/Latino, 11.11% Multiracial, and 1.8% other specified race. We used recruitment quotas to screen for a specified distribution of weekly alcohol use: 25% of the sample reported drinking 1-4 drinks a week, 40% drank 5-9 drinks a week, 25% drank 10-13 drinks a week, and 10% drank 14+ drinks a week. We did not recruit specifically for cannabis use, but 91. 2% of the participants reported lifetime cannabis use, and of these, 78. 9% had used in the past year and 60% in the past month.

Procedure

Participants completed a questionnaire battery and a series of free response questions. In this paper, our focus is on three free-response questions (denoted as FR). Free-response questions FR1 preceded FR2 and all three free-response preceded the questionnaire battery.

FR1 (Alcohol) Participants were asked: *"What are your beliefs about the potential benefits and harms of using alcohol?"* This substance-specific question was designed to elicit both positive and negative attributes that come to mind for each participant and are likely to be relevant to their understanding of addiction.

FR1 (Cannabis) Participants were asked: *"What are your beliefs about the potential benefits and harms of using cannabis?"*. This substance-specific question is directly analogous to FR1 Alcohol.

FR2: What is addiction? Participants were asked: *"Explain to me like I'm five, what is addiction?"* We chose this phrasing to elicit simple responses in a form that is likely to be familiar to our US-based participants. The open-ended nature of the question was designed so that participants were free to focus on whichever attributes of the concept came to mind most easily. We were concerned that an overly literal interpretation of *like I'm five* might prevent participants from including attributes that they perceive to be inappropriate for a younger audience, but we judged this limitation to be outweighed by its benefits. In particular, we anticipated that this framing might encourage people to share basic aspects of their understanding that they might otherwise omit in the context of a formal study because they are thought to be common knowledge.

Feature Extraction using GPT4o

Extraction pipeline To extract structured linguistic insights from participants' answers to the free response questions, we used ChatGPT (GPT4o). Each individual free-response was analyzed separately. Our prompt provided the model with the (1) DSM-5 section on Substance Use Disorders (SUD) and, (2) depending on the substance of interest, either the Alcohol Use Disorders or Cannabis Use Disorders DSM-5 subsection. The prompt instructed ChatGPT to evaluate to what degree a specific diagnostic feature was represented in the free response passage. A rating was obtained on 0 (definitely not represented) to 10 (definitely represented) scale. A separate call to the model's API was made for each of the 11 DSM-5 SUD diagnostic attributes (to ensure independence of ratings). Of relevance to non-clinical readers, the 11 symptoms of Alcohol Use Disorder and Cannabis Use Disorder are identical in nature except with different target substance. Table 1 enumerates the 11 DSM-5 SUD diagnostic attributes that we used as structured feature representations.

FR2 substance-specific context manipulation The free-response question FR2 *"Explain to me like I'm five, what is addiction?"* was not substance-specific during data collection, in that the question did not ask participants to explain addiction in relation to either alcohol or cannabis specifically. However, we obtained from ChatGPT two separate annotations for each participant's response to FR2. One annotation using the version of our prompt that contextualized the response using the Alcohol Use Disorders DSM-5 subsection (henceforth: FR2 Alcohol) and one annotation using the version of our prompt that contextualized the response using the Cannabis Use Disorders DSM-5 subsection (henceforth: FR2 Cannabis). While these subsections state effectively the same diagnostic criteria, here they create a substance-specific prompt context in which the model is asked to provide a response. This allowed us to examine biases that arise in response annotation directly as a consequence of this substance-specific context, holding the content of the participant's response constant. To evaluate the test-retest reliability of GPT4o's ratings, this process was repeated 10 times for each FR2 Alcohol and FR2 Cannabis. Intraclass correlations were used as the index of this reliability estimate. For the subsequent analyses, the means of these symptoms across the 10 runs were used.

Table 1. Diagnostic Feature DSM-5 Language

Feature Label	Description
Excessive Consumption	(Alcohol/Cannabis) is often taken in larger amounts or over a longer period than was intended.
Unsuccessful Cutting	There is a persistent desire or unsuccessful efforts to cut down or control (alcohol/cannabis) use.
Excessive Time	A great deal of time is spent in activities necessary to obtain (alcohol/cannabis), use (alcohol/cannabis), or recovering from its effects.
Craving	Craving, or a strong desire or urge to use (alcohol/cannabis).
Role Obligations	Recurrent (alcohol/cannabis) use resulting in a failure to fulfill major role obligations at work, school, or home.
Interpersonal	Continued (alcohol/cannabis) use despite having persistent or recurrent social or interpersonal problems caused or exacerbated by the effects of alcohol.
Activities	Important social, occupational, or recreational activities are given up or reduced because of (alcohol/cannabis) use.
Hazardous Use	Recurrent (alcohol/cannabis) use in situations in which it is physically hazardous.
Persistent Problems	(Alcohol/Cannabis) use is continued despite knowledge of having a persistent or recurrent physical or psychological problem that is likely to have been caused or exacerbated by (alcohol/cannabis).
Tolerance	1. A need for markedly increased amounts of (alcohol/cannabis) to achieve intoxication or desired effect. OR 2. A markedly diminished effect with continued use of the same amount of (alcohol/cannabis).
Withdrawal	1. The characteristic withdrawal syndrome for (alcohol/cannabis). OR 2. (Alcohol/cannabis) (or a closely related substance) is taken to relieve or avoid withdrawal symptoms.

Data Analytic Plan

Aims Our first aim was to examine which features were represented most frequently in participants' responses, and to examine any relationships between features. We also aimed to explore whether there was evidence for substance-specific biases in participants' beliefs around the benefits and harms of alcohol and cannabis, operationalized in terms of differences in the extent to which the 11 diagnostic features were represented in responses to FR1 Alcohol and FR1 Cannabis. Finally, we aimed to explore whether there was evidence for substance-specific biases in ChatGPT's representations of FR2, operationalized in terms of the differences in the extent to which diagnostic features were judged to be represented in FR2 when presented in the context of Cannabis (FR2 Cannabis) and Alcohol (FR2 Alcohol).

Paired differences in feature representation Paired samples t-tests were run for each one of the 11 diagnostic features to characterize the differences in rate of representation between the substances for each question.

Relationships between features We also aimed to ascertain the structure of symptom covariation present in participants' free-responses and in ChatGPT's substance-specific annotations of FR2 (e.g., to the degree that a response contains relevance to craving, to what degree is withdrawal also present). To do this, we calculated correlations between the representation levels of diagnostic symptoms within responses for each of three free-response questions.

Alignment of substance-specific biases between participants and ChatGPT We aimed to examine whether any substance specific biases evidenced in human responses to FR1 Alcohol and FR1 Cannabis were similarly evidenced in ChatGPT's substance-specific annotations of FR2. To examine this, we first calculated the *differences* between the diagnostic-feature representation rates in ChatGPT's substance-specific annotations of FR2 (e.g. did ChatGPT detect the *Persistent Use* feature more frequently in FR2 Alcohol than FR2 Cannabis; the difference in detection rates quantifies this bias). We then calculated analogous *differences* in human participant responses to FR1 Alcohol and FR1 Cannabis. We were able to assess statistical evidence for systematic differences in substance-specific bias by testing for differences between these distributions of differences. Again, we conducted paired samples t-tests between the *differences* of FR2 Alcohol and FR2 Cannabis with the *differences* of FR1 Alcohol and FR1 Cannabis for each diagnostic symptom. Finally, we also examined correlations between difference scores for FR2 and for FR1.

Together, these analyses allowed us to quantify the structure of participants' beliefs around the benefits and harms of alcohol and cannabis, the attributes of their intuitive explanations of the concept of addiction, and any systematic substance-specific biases conveyed by participants and by the ChatGPT annotation procedure itself.

Table 2. Example Question Responses and Their Single Run ChatGPT Ratings

Example Response	"Addiction is when your body really needs something or else you feel sick and feel like you can't do anything, even though the thing you want is bad for you."		"I think it's a bit of a net negative in general because it feels like it's very easy to get addicted to it. However, it's something that's a lot more exciting than a regular beverage to share with friends so it isn't all bad."	"I think cannabis is amazing for also boosting confidence and keeping your mood controlled. I think one harm is it kills productiveness, makes you a bit lazy, and isnt healthy for your lungs."
	FR2 Alcohol	**FR2 Cannabis**	**FR1 Alcohol**	**FR1 Cannabis**
Excessive Consumption	2	3	2	4
Unsuccessful Cutting	2	3	2	2
Excessive Time	3	2	0	3
Craving	3	2	3	4
Role Obligations	2	0	2	3
Interpersonal	3	2	2	2
Activities	2	0	2	3
Hazardous Use	3	3	2	2
Persistent Use	2	0	3	3
Tolerance	3	4	0	2
Withdrawal	2	2	0	0

Results

FR1: Benefits and harms of alcohol and cannabis

We examined the structure of participant's beliefs around the benefits and harms of alcohol and cannabis, first exploring substance-specific biases in belief attributes, and then examining the relationships between belief attributes. Figure 1 Panel A shows the overall distribution of feature representation (diagonal) and feature correlation (upper and lower triangles) between diagnostic features for both alcohol (FR1 Alcohol, yellow) and cannabis (FR1 Cannabis, green).

Substance-specific biases in participants' beliefs Paired samples t-tests identified significant differences in the beliefs that participants conveyed in response to FR1 Alcohol and FR1 Cannabis for 5 out of the 11 diagnostic features ($p < .005$). Specifically, participants' beliefs around the benefits and harms of alcohol showed statistically significantly higher representation for the diagnostic criteria: *Excessive Consumption* ($t(141) = 3.17$, $p = .002$, Cohen's $d = .27$), *Interpersonal* ($t(141) = 4.73$, $p < .001$, $d = .40$), *Hazardous Use* ($t(141) = 2.42$, $p < .001$, $d = .20$), and *Persistent Use* ($t(141) = 3.10$, $p = .003$, $d = .26$). In contrast, participants' beliefs around the benefits and harms of cannabis showed statistically significantly higher representation of *Withdrawal* ($t(141) = -3.12$, $p = .002$, $d = -.26$) than did their beliefs around alcohol. One way to interpret these differences is that overall, participant's beliefs around the benefits and harms of alcohol include increased representation of psychopathology relative to cannabis, in line with prior survey results: alcohol is believed to be more problematic than cannabis. Excess Time ($t(141) = 1.93$, $p = .06$) and Tolerance ($t(141) = -1.93$, $p = .07$) fell beyond the $p < .05$ alpha level, but showed some notable effect of varying beliefs around alcohol and cannabis. The remaining comparisons did not provide statistically significant evidence in our dataset of substance-specific biases in these features.

Relational structure in participants' beliefs We calculated pairwise correlations between the representation of diagnostic features in people's beliefs around alcohol (FR1 Alcohol) and cannabis (FR1 Cannabis) within-substance. Among diagnostic features represented in people's beliefs around alcohol the mean correlation was $r = .52$, however the results ranged from $r = .23$ (FR1 Alcohol *Persistent Use*, FR1 Alcohol *Excessive Use*) to $r = .79$ (FR1 Alcohol *Interpersonal*, FR1 Alcohol *Role Obligations*). Overall, there was a relatively high degree of covariation between diagnostic features among people's belief about alcohol. We observed similar patterns of correlations between features among people's beliefs around the benefits and harms of cannabis (average pairwise correlation was $r = .54$). The most strongly correlated features of people's beliefs about cannabis were *Activities* and *Role Obligations* ($r = .79$), for example. Compared to people's beliefs about alcohol, the mean correlations within cannabis for each diagnostic feature were similarly high.

FR2: What is addiction?

We examined the structure of participant responses to FR2 – *"Explain to me like I'm five, what is addiction?"* Recall that we analyzed each participant's response twice: once using a prompt that contextualized the response as relevant to cannabis (FR2 Cannabis) and once using a

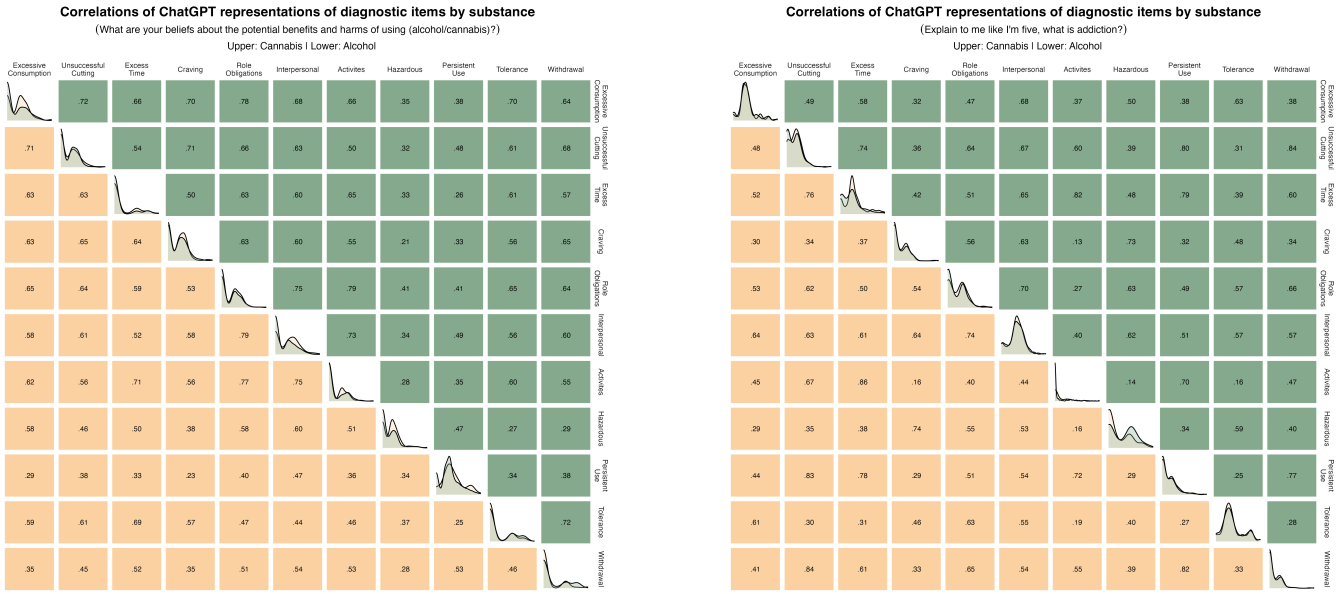

Figure 1: Conceptual Structure of Substance Use Disorders for Humans (panel A) and GPT4o (panel B)

prompt that contextualized the response as relevant to alcohol (FR2 Alcohol). This manipulation allowed us to treat these two feature representations as if they were distinct substance-specific responses, and therefore replicate the analyses we reported above on FR1 Cannabis and FR1 Alcohol. This offered a way to assess the potential for substance-specific biases generated by the *ChatGPT annotation pipeline itself*. In principle, there should be minimal differences in feature distribution between these two representations, because the responses provided to the model in FR2 Alcohol and FR2 Cannabis are the same responses. Any significant differences therefore indicate algorithmic bias. Intraclass correlations between the 10 runs showed good and equal test-retest reliability for both FR2 Alcohol (mean ICC = .86, range = .78 to .94) and FR2 Cannabis (mean ICC = .86, range = .78 to .93).

Substance-specific biases in LLM annotation A series of 11 paired samples t-tests indicated significant differences in feature representation between FR2 Alcohol and FR2 Cannabis for 7 different diagnostic features: *Unsuccessful Cutting* ($t(148) = 6.47$, $p < .001$, Cohen's d = .53), *Craving* ($t(148) = 2.12$, $p = .04$, d = .17), *Role Obligations* ($t(148) = 6.52$, $p < .001$, d = .53), *Activities* ($t(148) = 3.81$, $p < .001$, d = -.08), *Hazardous Use* ($t(148) = -9.13$, $p < .001$, d = -.75), *Persistent Use* ($t(148) = 3.69$, $p < .001$, d = .30), and *Excess Time* ($t(148) = 10.77$, $p < .001$, d = .88). Of these significant results, all of the diagnostic features except *Hazardous Use* were judged to be more strongly represented in participants' explanations of addiction when those explanations were contextualized as relevant to alcohol (FR2 Alcohol) than when they were contextualized as relevant to cannabis (FR2 Cannabis). Recall that all features considered here are *diagnostic* features related to SUDs. In that light, any systematically increased substance-specific detection of features indicates a systematic algorithmic bias towards a problematic or clinically sensitive interpretation of language surrounding the substance. These results suggest that ChatGPT-4o has a greater probability of interpreting language around addiction as evidencing clinically relevant diagnostic symptoms when contextualized as relevant to alcohol than when contextualized as relevant to cannabis.

Relational structure in participants' explanations of addiction We examined whether there were meaningful relationships between the features participants mentioned in their explanations of addiction. We performed two pairwise correlation analyses. The first examined correlations between diagnostic features represented in FR2 Alcohol. In this analysis, the *average* pairwise correlation between features was $r = .52$. Results ranged from r= .20 (FR2 Alcohol *Activities*, FR2 Alcohol *Hazardous Use*) to $r = .85$ (FR2 Alcohol *Activities*, FR2 Alcohol *Excess Time*). The second analysis examined correlations between feature in FR2 Cannabis. The mean pairwise correlation here was $r = .49$, ranging from $r = .08$ (FR2 Cannabis *Activities*, FR2 Cannabis *Craving*) to $r = .83$ (FR2 Cannabis *Activities*, FR2 Cannabis *Excess Time*). Together, our analyses of the representation of diagnostic features (and their correlations) in participants' explanations of addiction (FR2 Alcohol, FR2 Cannabis) reveal that GPT4o exhibits a bias that attributes clinically-relevant diagnostic more readily in the context of alcohol than cannabis.

Differences between LLM and Human biases

We performed a comparison between substance-specific biases exhibited by participant responses (FR1 Alcohol and

FR1 Cannabis) and those generated by ChatGPT (FR2 Alcohol and FR2 Cannabis), using a series of pairwise t-tests. First, we computed *differences* in feature representation between the two substances for FR1 (e.g. did participants mention *Craving* more frequently when describing the benefits and harms of alcohol than cannabis?) and the analogous differences in FR2 (e.g. did ChatGPT attribute *Craving* more frequently to explanations of addiction when contextualized as relevant alcohol versus cannabis?). We then computed the *differences in the differences*. This allowed us to ask whether substance-specific biases present in participants' substance-specific responses (FR1) were aligned with the biases attributable to ChatGPT (FR2).

We found significant differences related to Excessive Consumption (t(122) = -3.50, $p < .001$, Cohen's d = -.32), Interpersonal (t(122) = -5.41, $p < .001$, d = -.49), Hazardous Use (t(122) = -4.46, $p < .001$, d = -.39), Persistent Use (t(122) = -2.26, p = .03, d = -.20), Excess Time (t(122) = 2.29, p = .02, d = .21), and Withdrawal (t(122) = -2.26, p = .03, d = -.20). All of these, with the exception of Excess Time, showed significantly stronger bias towards alcohol in FR1 than in FR2. These results imply that alcohol-specific biases were stronger in human responses than the LLM biases, which sometimes favored cannabis over alcohol. At least 5 of 11 features showed greater alcohol-specific bias in participants responses (FR1 Alcohol and FR1 Cannabis). Only 1 diagnostic feature was associated with a stronger alcohol-specific bias in FR2 than in FR1.

Finally, we calculated the pairwise correlations between all the bias metrics (alcohol minus cannabis) for FR2 and FR1 and found meaningful differences. The mean pairwise correlation between bias metrics (difference scores between alcohol versus cannabis) per symptom for FR2 was $r = .16$ whereas the mean FR1 pairwise correlation was $r = .37$. The range was similarly disparate with FR2 ranging from $r = -.03$ (Activities, Withdrawal) to $r = .37$ (Interpersonal, Unsuccessful Cutting) while FR1 ranged from $r = .01$ (Hazardous Use, Withdrawal) to $r = .68$ (Unsuccessful Cutting, Craving). As a whole, the analyses comparing the differences in alcohol and cannabis for FR1 and FR2 suggest that there are meaningful differences in the structure of representations between ChatGPT and humans biases between substances.

Discussion

Participants were asked to share their beliefs around the benefits and harms of alcohol and cannabis, and to explain simply what it means to be *addicted*. Using a frontier language model (ChatGPT-4o) we extracted structured representations of participant's explanations. We used this pipeline to evaluate the extent to which 11 clinically relevant diagnostic symptoms (DSM-5 section on Substance Use Disorders; Alcohol Use Disorders & Cannabis Use Disorders subsections) were represented in participant's beliefs and explanations. A series of analyses assessing differences in the extracted representations revealed evidence for systematic, substance-specific biases in human beliefs.

People described more diagnostic features when talking about alcohol. 4 of the 11 diagnostic features were statistically overrepresented in beliefs about alcohol relative to cannabis (compared to only 1 of the 11 features being over-represented when talking about cannabis). Correlation analyses showed meaningful correlations between the diagnostic symptoms people mentioned in their beliefs about both substances (mean FR1 Alcohol $r = .52$, mean FR1 Cannabis $r = .54$).

To understand if ChatGPT-4o itself contributes substance-specific biases, we performed a context manipulation on people's explanations of addiction (which were not substance-specific). We extracted structured representations from explanations contextualized by alcohol (FR2 Alcohol) vs cannabis (FR2 Cannabis) using the respective DSM-5 sections. We repeated this process 10 times to establish robust bias estimates. ChatGPT attributed greater representation of the diagnostic features to explanations when they were contextualized as relevant to alcohol (6 out of 11 features; only over-attributing 1 feature to cannabis).

We also examined the differences between alcohol and cannabis biases for both FF1 and FF2. We found that ChatGPT alcohol-specific biases were only stronger than human alcohol-specific biases for 1 of the 11 features; human alcohol-specific bias were meaningfully higher for 5 out of 11. Participants' beliefs show more systematic biases than those imposed by ChatGPT. Our results are well-aligned with existing literature on LLMs and their potential inherent biases from their training data which can lead to skewed outputs particularly around sensitive concepts such as addiction (Timmons et al., 2023; Spallek et al., 2023; Soun & Nair, 2023) and with known human biases towards alcohol being more related to harms and addiction issues, a conclusion heavily supported in the literature (Allen et al., 2018). However, there is evidence that cannabis use, in particular chronic use, can produce cognitive impairments (Copeland, Clement, & Swift, 2014; Frolli et al., 2021; Lichenstein, Shaw, & Forbes, 2022) and has been linked to a range of mental health concerns beyond CUD (Choi, DiNitto, Marti, & Choi, 2016; Copeland et al., 2014; DiNitto & Choi, 2011; Hall, 2009; Tibbo et al., 2018). The results of our study support the importance of a deeper understanding of potential biases in ChatGPT's conceptualization of addiction and other clinically relevant topics, which may be significantly different from human perceptions in some diagnostic features. Future work is also needed to analyze these biases more systematically and explore other areas of clinical concepts beyond the DSM-5. Finally, our findings shed light on how biases in conceptual structure may be ingrained in language models as a consequence of their training, yet the resulting biases may differ from what is seen in clinical populations themselves.

References

Ahn, J., & Oh, A. (2021). Mitigating language-dependent ethnic bias in bert. *arXiv preprint arXiv:2109.05704*.

Allen, J. A., Farrelly, M. C., Duke, J. C., Kamyab, K., Nonnemaker, J. M., Wylie, S., … Gourdet, C. (2018). Perceptions of the relative harmfulness of marijuana and alcohol among adults in oregon. *Preventive Medicine*, *109*, 34–38.

Bordia, S., & Bowman, S. R. (2019). Identifying and reducing gender bias in word-level language models. *arXiv preprint arXiv:1904.03035*.

Choi, N. G., DiNitto, D. M., Marti, C. N., & Choi, B. Y. (2016). Relationship between marijuana and other illicit drug use and depression/suicidal thoughts among late middle-aged and older adults. *International Psychogeriatrics*, *28*(4), 577–589.

Clark, M. (2011). Conceptualising addiction: How useful is the construct. *International Journal of Humanities and Social Science*, *1*(13), 55–64.

Copeland, J., Clement, N., & Swift, W. (2014). Cannabis use, harms and the management of cannabis use disorder. *Neuropsychiatry*, *4*(1), 55–63.

Degenhardt, L., Charlson, F., Ferrari, A., Santomauro, D., Erskine, H., Mantilla-Herrara, A., … others (2018). The global burden of disease attributable to alcohol and drug use in 195 countries and territories, 1990–2016: a systematic analysis for the global burden of disease study 2016. *The Lancet Psychiatry*, *5*(12), 987–1012.

DiNitto, D. M., & Choi, N. G. (2011). Marijuana use among older adults in the usa: user characteristics, patterns of use, and implications for intervention. *International Psychogeriatrics*, *23*(5), 732–741.

Douaihy, A., & Daley, D. (2013). *Substance use disorders*. Oxford University Press.

Frolli, A., Ricci, M. C., Cavallaro, A., Lombardi, A., Bosco, A., Di Carmine, F., … Franzese, L. (2021). Cognitive development and cannabis use in adolescents. *Behavioral Sciences*, *11*(3), 37.

Hall, W. (2009). The adverse health effects of cannabis use: what are they, and what are their implications for policy? *International Journal of drug policy*, *20*(6), 458–466.

Hu, T., Kyrychenko, Y., Rathje, S., Collier, N., van der Linden, S., & Roozenbeek, J. (2024). Generative language models exhibit social identity biases. *Nature Computational Science*, 1–11.

Leung, J., Chan, G. C., Hides, L., & Hall, W. D. (2020). What is the prevalence and risk of cannabis use disorders among people who use cannabis? a systematic review and meta-analysis. *Addictive behaviors*, *109*, 106479.

Lichenstein, S. D., Shaw, D. S., & Forbes, E. E. (2022). Cannabis, connectivity, and coming of age: Associations between cannabis use and anterior cingulate cortex connectivity during the transition to adulthood. *Frontiers in Human Neuroscience*, *16*, 951204.

Lu, W., Lopez-Castro, T., & Vu, T. (2023). Population-based examination of substance use disorders and treatment use among us young adults in the national survey on drug use and health, 2011–2019. *Drug and Alcohol Dependence Reports*, *8*, 100181.

Morales, S., Clarisó, R., & Cabot, J. (2024). A framework to model ml engineering processes. *arXiv preprint arXiv:2404.18531*.

Musolesi, M. (2024). Creative beam search: Llm-as-a-judge for improving response generation..

O'neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Ranjan, R., Gupta, S., & Singh, S. N. (2024). A comprehensive survey of bias in llms: Current landscape and future directions. *arXiv preprint arXiv:2409.16430*.

Soun, R. S., & Nair, A. (2023). Chatgpt for mental health applications: A study on biases. In *Proceedings of the third international conference on ai-ml systems* (pp. 1–5).

Spallek, S., Birrell, L., Kershaw, S., Devine, E. K., Thornton, L., et al. (2023). Can we use chatgpt for mental health and substance use education? examining its quality and potential harms. *JMIR Medical Education*, *9*(1), e51243.

Tibbo, P., Crocker, C. E., Lam, R. W., Meyer, J., Sareen, J., & Aitchison, K. J. (2018). Implications of cannabis legalization on youth and young adults. *The Canadian Journal of Psychiatry*, *63*(1), 65–71.

Timmons, A. C., Duong, J. B., Simo Fiallo, N., Lee, T., Vo, H. P. Q., Ahle, M. W., … Chaspari, T. (2023). A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. *Perspectives on Psychological Science*, *18*(5), 1062–1096.