Building Healthy Recommendation Sequences for Everyone:
A Safe Reinforcement Learning Approach

FAccTRec Workshop
ACM RecSys 2020

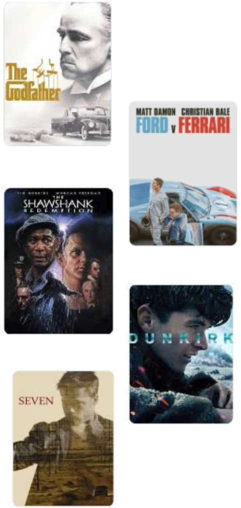
Ashudeep Singh, Yoni Halpern, Nithum Thain, Konstantina Christakopoulou, Ed H. Chi, Jilin Chen, and Alex Beutel

Cornell University Google AI

Building **Healthy** Recommendation Sequences for **Everyone**

How should a recommender system responsibly make item recommendations when selecting from a corpus that contains **some amount of potentially unhealthy content** like violence in movies or poor nutrition in restaurants?

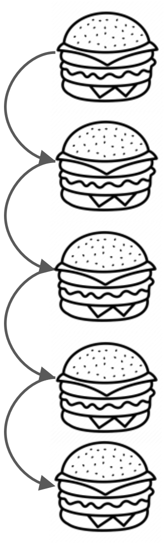
What is an unhealthy user experience? (Hypothetical Ex.)



User Trajectory τ_1	Violence score		User Trajectory τ_2	Violence score
	0.7	VS		0.7
	0.85			0.2
	0.75			0.33
	0.67			0.5
	0.8			0.67
Average Violence-score experienced by the user:	0.75			0.40

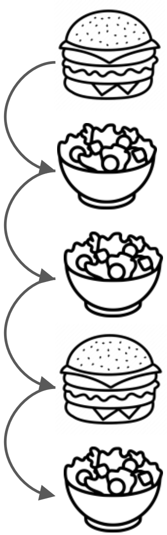
What is an unhealthy user experience? (Hypothetical Ex.)

User Trajectory τ_1



Happiness	Health Risk
0.9	0.9
0.9	0.9
0.9	0.9
0.9	0.9
0.9	0.9
0.9	0.9

User Trajectory τ_2

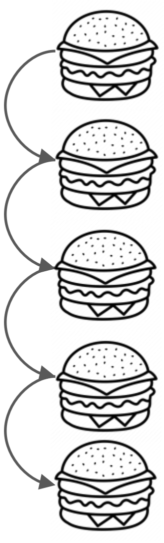


Happiness	Health Risk
0.9	0.9
0.6	0.1
0.6	0.1
0.9	0.9
0.6	0.1
0.72	0.42

Average unhealthiness of food recs experienced by the user.

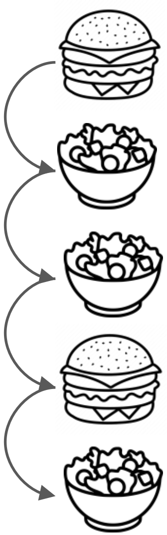
What is an unhealthy user experience? (Hypothetical Ex.)

User Trajectory τ_1



Happiness	Health Risk
0.9	0.9
0.9	0.9
0.9	0.9
0.9	0.9
0.9	0.9
0.9	0.9

User Trajectory τ_2



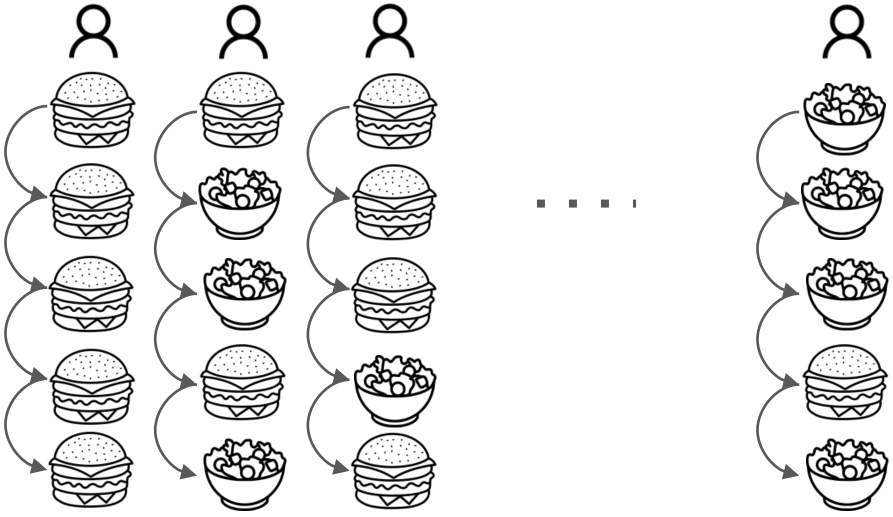
Happiness	Health Risk
0.9	0.9
0.6	0.1
0.6	0.1
0.9	0.9
0.6	0.1
0.72	0.42

Average unhealthiness of food recs experienced by the user.

Aggregate the health risk as the average of the healthiness scores (e.g. Violence etc.) over the sequence of recommendations.

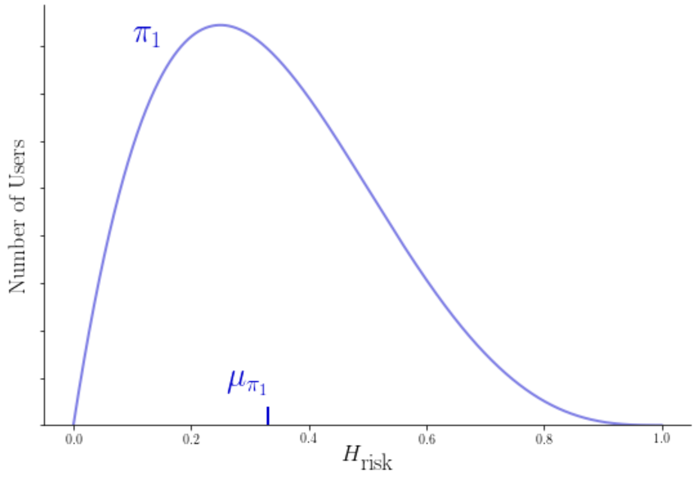
In general, we can aggregate it using any other function of the health risk of the trajectory.

What is an unhealthy user experience?

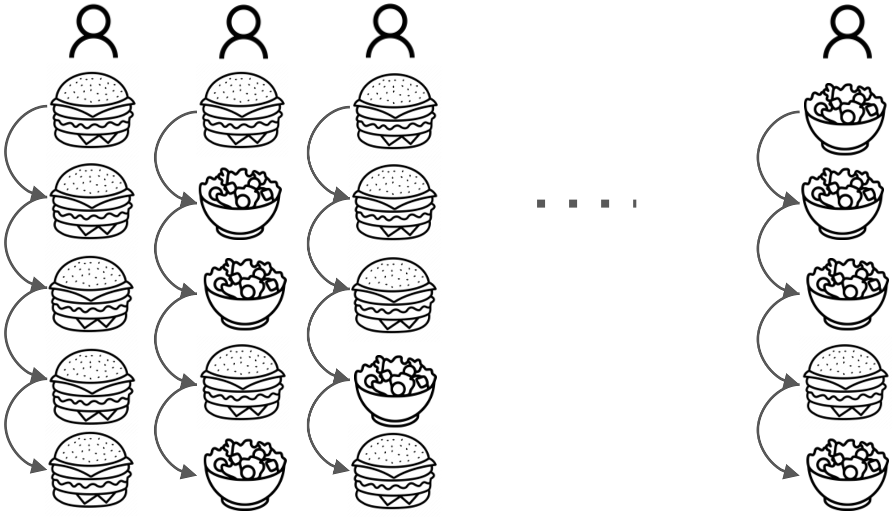


Average Health Risk 0.9 0.42 0.74 . . . . 0.26

For a recommendation policy $\boldsymbol{\pi}_{1,}$ consider the distribution of health risk over the set of users.



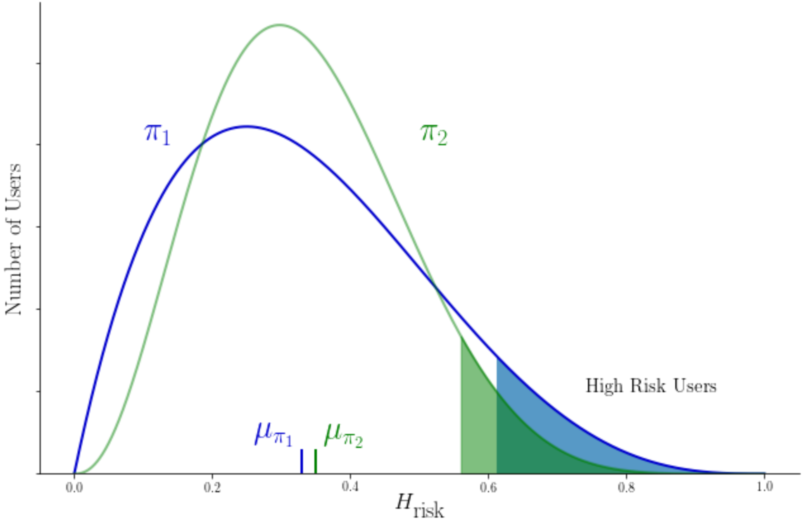
What is an unhealthy user experience?



Average Health Risk 0.9 0.42 0.74 .... 0.26

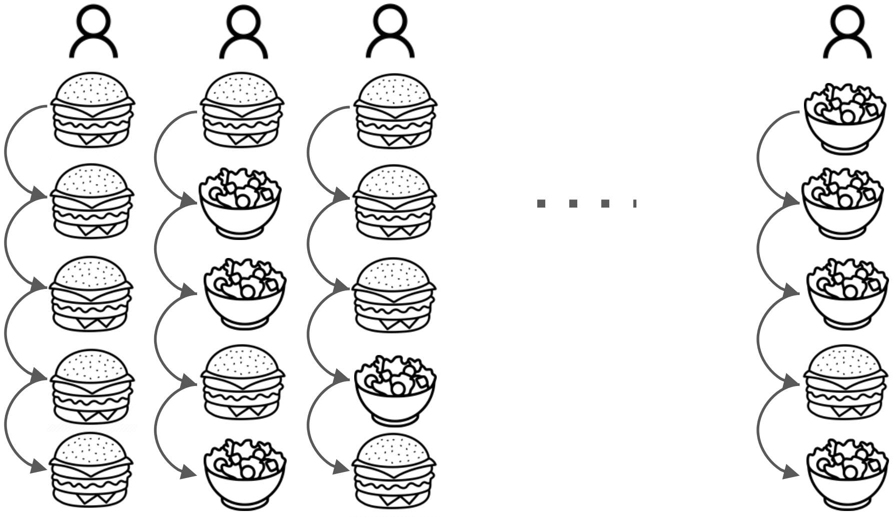
For a recommendation policy $\boldsymbol{\pi}_{1,}$ consider the distribution of health risk over the set of users.

Compare it to the policy $\boldsymbol{\pi}_2$.



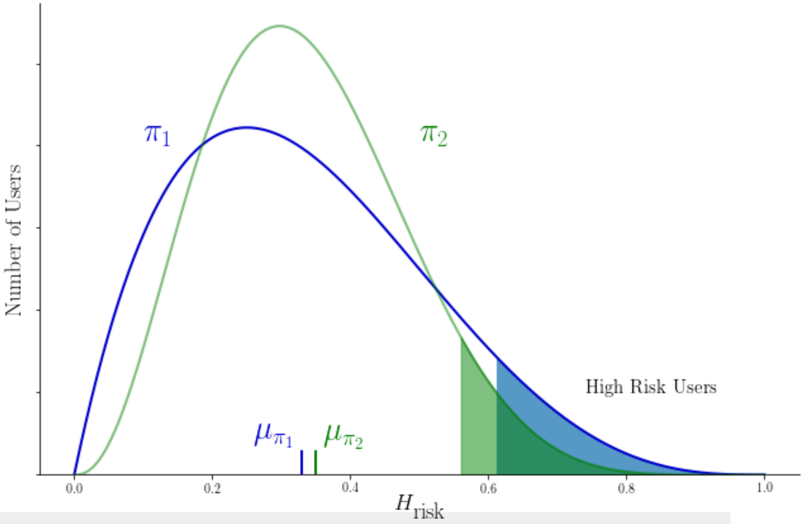
Which out of $\boldsymbol{\pi}_1$and $\boldsymbol{\pi}_2$ is preferable?

What is an unhealthy user experience?



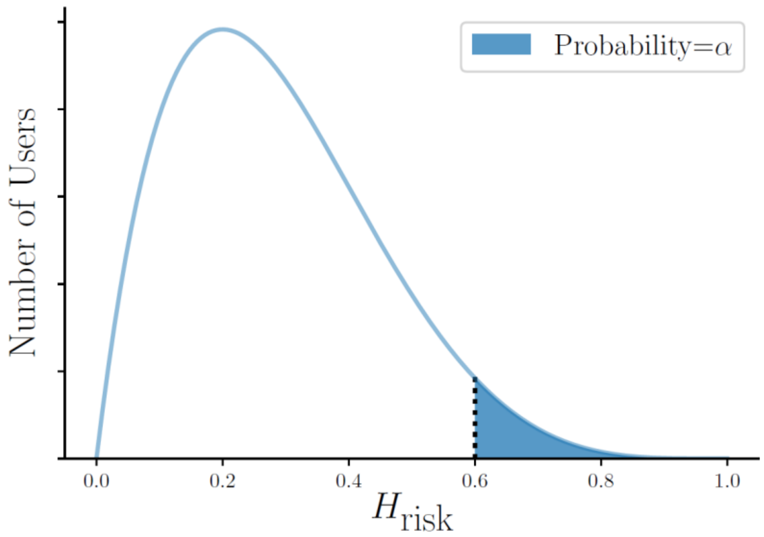
Average Health Risk 0.9 0.42 0.74 . . . . 0.26

Consider the distribution of health risks over the worst-case user trajectories.



Safety concern: User trajectories with very high health risks.

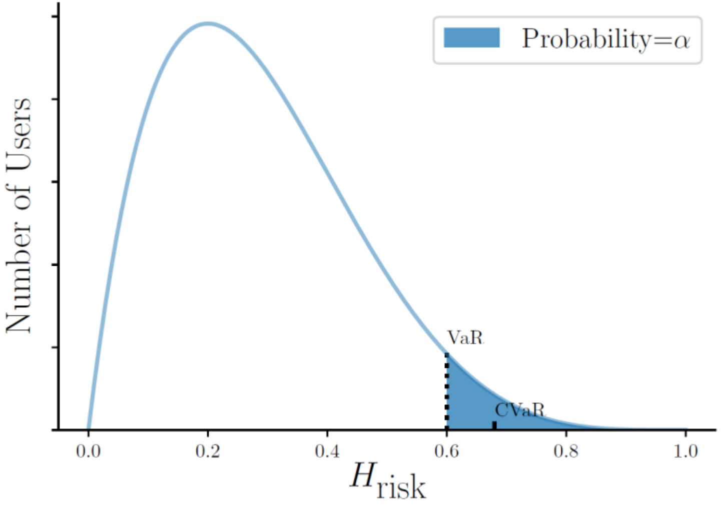
Safety Goal



- Minimize the risk for the worst-case users.
- How to quantify the risk to the worst-case users?

Answer: Conditional Value-at-Risk

Risk Measure: Conditional Value-at-Risk (CVaR)



CVAR$_\alpha$= Expectation of health risk in the shaded region.

- VaR$_\alpha$ is defined as 1-α^{th} percentile risk.
- CVaR$_\alpha$ is the expected value of risk below VaR$_\alpha$.

$$\text{CVAR}_\alpha = \mathbb{E}[\ H_{\text{risk}} \mid H_{\text{risk}} \geq \text{VaR}_\alpha\ ]$$

Our overall objective is to **maximize User Satisfaction** while **minimizing worst-case health risk (CVaR)**.

Key contributions

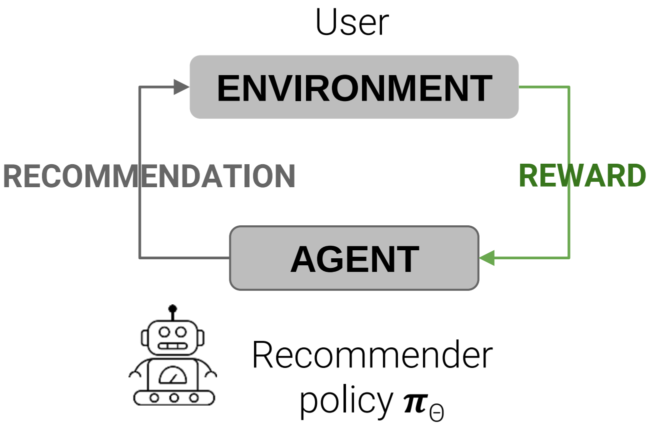
- Propose definition for safety in recommender systems.

- Propose a Safe-RL technique for training recommender systems for sequential recommendations with multiple objectives.

- Demonstrate the effectiveness of safety-constrained policies on a simulated experiment.

Problem Setup

We frame the problem of **sequential recommendations as a RL problem**.



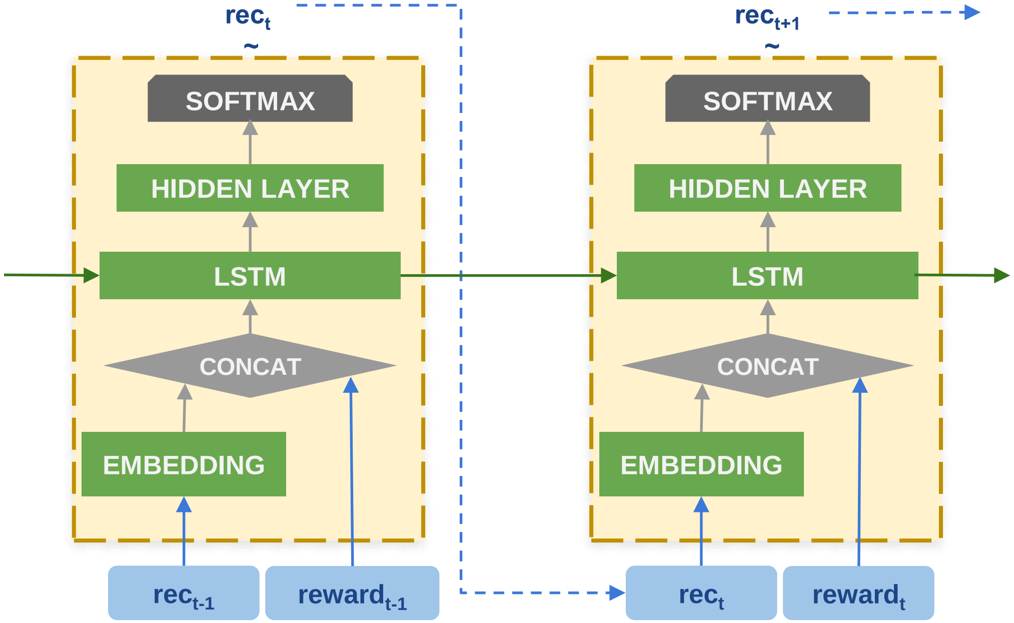
- Each experimental setup consists of an **environment** and an **agent**.
- The goal is to learn an agent that adapts by interacting with the environment without knowing the underlying dynamics.

Agent: RNN Recommender for Sequential Recs

We use an RNN recommender agent similar to Chen et al. (2019).

At t=0, a user comes in with a hidden preference vector, and the agent is given as input (<s_0>, 0) where <s_0> indicates a starting token.

In the following steps t=1,2,...T, the agent samples a recommendation from the softmax (with no repetitions allowed).



Objective: Safe Sequential Recommendation

- Maximize Rewards from the users (CTR, happiness, etc.), and

- Minimize worst-case health risk (CVaR).

$$\boldsymbol{\pi}_{\boldsymbol{\theta}}{}^{*} = \mathrm{argmax}_{\boldsymbol{\pi}}\, \mathbb{E}_{\tau\sim\boldsymbol{\pi}}[\mathrm{R}(\tau)] - \lambda\, \mathrm{CVaR}_{\boldsymbol{\alpha}}(\mathrm{H}_{\mathrm{risk}}|\boldsymbol{\pi}).$$

Safe Reinforcement Learning Objective

Optimal Policy

Total Reward of the Trajectory

α: Cutoff percentile

Optimizing for the Safety Constraint

Following Rockafellar and Uryasev (2000), CVaR can be expressed as :

$$\mathrm{CVaR}_{\alpha}(\mathrm{H}_{\mathrm{risk}}|\pi) = \left[\nu_{\alpha}(\pi) + \frac{1}{1-\alpha}\mathbb{E}_{\pi}[(\mathrm{H}_{\mathrm{risk}}(\tau) - \nu_{\alpha}(\pi))^{+}]\right]$$

Expectation of Risk beyond **ν**

$$\pi_{\theta^*} = \underset{\theta}{\operatorname{argmax}}\ \mathbb{E}_{\tau\sim\pi_{\theta}}\left[R(\tau) - \lambda\left(\nu_{\alpha}(\pi) + \frac{1}{1-\alpha}(\mathrm{H}_{\mathrm{risk}}(\tau) - \nu_{\alpha}(\pi))^{+}\right)\right]$$

Tamar et al. (2015): approximate the gradient by plugging in the VaR on the minibatch

$$\nabla_{\theta}\mathbb{E}_{\tau\sim\pi_{\theta}}R'(\tau, \nu_{\alpha}(\pi)) \approx \nabla_{\theta}\mathbb{E}_{\tau\sim\pi_{\theta}}R'(\tau, \tilde{\nu_{\alpha}})$$

Williams (1992): REINFORCE update

$$\nabla_{\theta}\mathbb{E}_{\tau\sim\pi_{\theta}}R'(\tau, \tilde{\nu_{\alpha}}) = \mathbb{E}_{\tau\sim\pi_{\theta}}\nabla_{\theta}\log P(\tau|\pi_{\theta})\left[R(\tau) - \frac{\lambda}{1-\alpha}(\mathrm{H}_{\mathrm{risk}}(\tau) - \tilde{\nu_{\alpha}})^{+}\right]$$

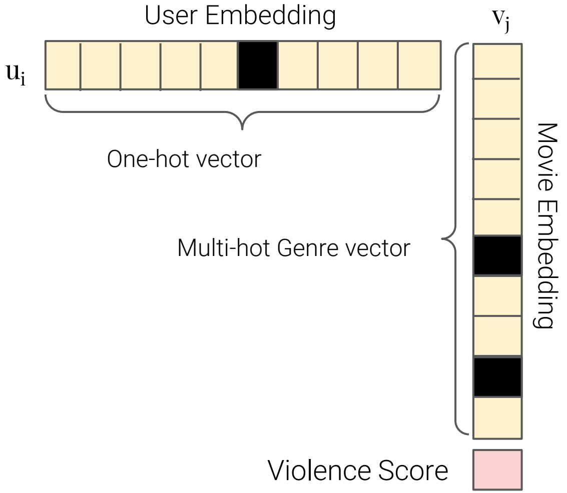
Agents

Reward Optimizing	$\boldsymbol{\pi}_{\boldsymbol{\theta}}{}^{*} = \mathrm{argmax}_{\boldsymbol{\pi}}\ \mathbb{E}_{\tau\sim\boldsymbol{\pi}}[R(\tau)]$
CVaR-Multiobjective (CVaR-MO)	$\boldsymbol{\pi}_{\boldsymbol{\theta}}{}^{*} = \mathrm{argmax}_{\boldsymbol{\pi}}\ \mathbb{E}_{\tau\sim\boldsymbol{\pi}}[R(\tau)] - \lambda\ \mathrm{CVaR}_{\boldsymbol{\alpha}}(H_{\mathrm{risk}}\vert\boldsymbol{\pi})$
Avg-Health Multiobjective (Avg-Health-MO)	$\boldsymbol{\pi}_{\boldsymbol{\theta}}{}^{*} = \mathrm{argmax}_{\boldsymbol{\pi}}\ \mathbb{E}_{\tau\sim\boldsymbol{\pi}}[R(\tau) - \lambda\ H_{\mathrm{risk}}(\tau)].$

Environment: Movie Recommendation Setup

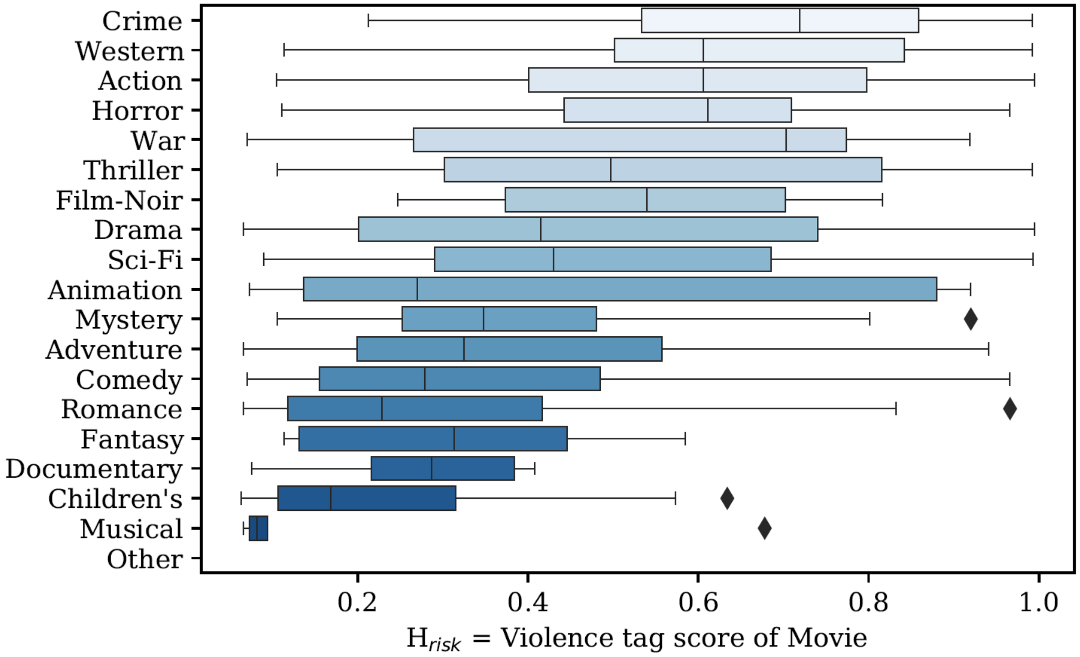
Each user prefers a particular genre i.e. rates every movie in that genre as 1, and everything else as 0.

$$r_{i,j} =$$



Idea: Some genres have a higher number of violent movies, and hence some users are vulnerable to more violent content under the reward optimal agent.

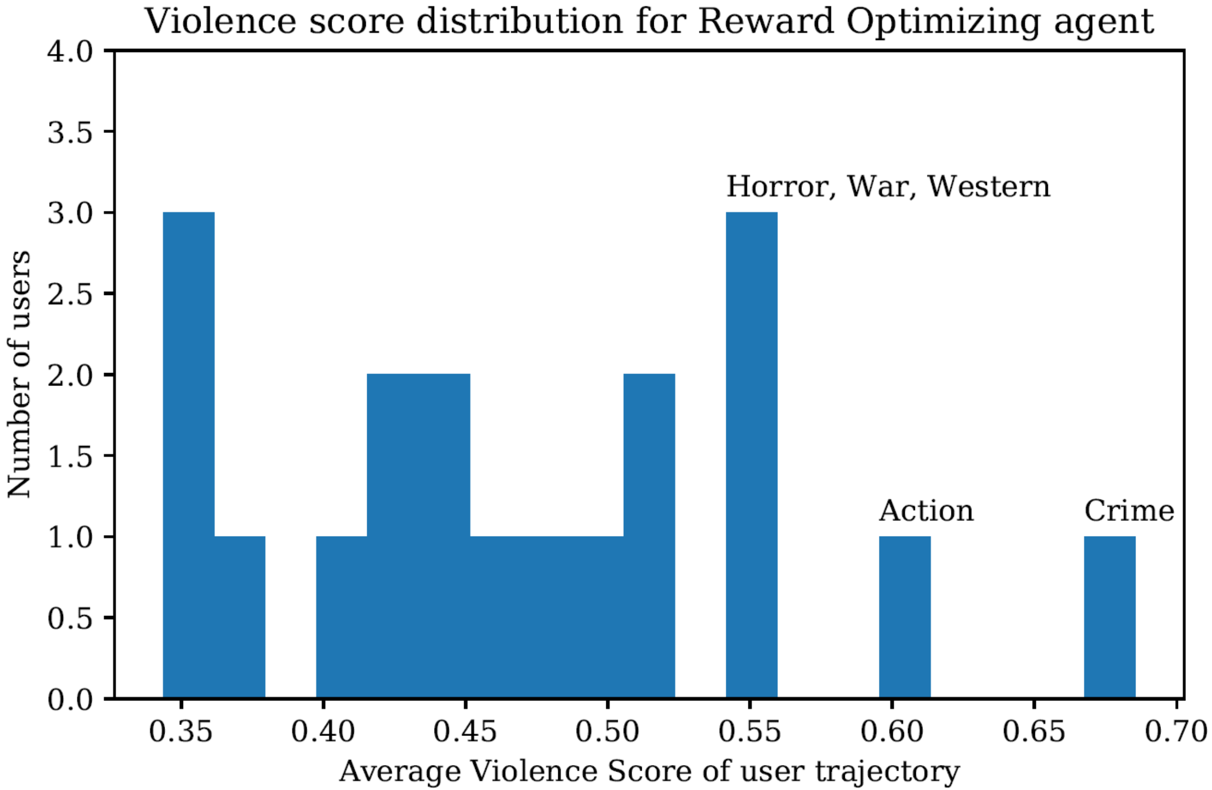
Distribution of violence scores for different genres



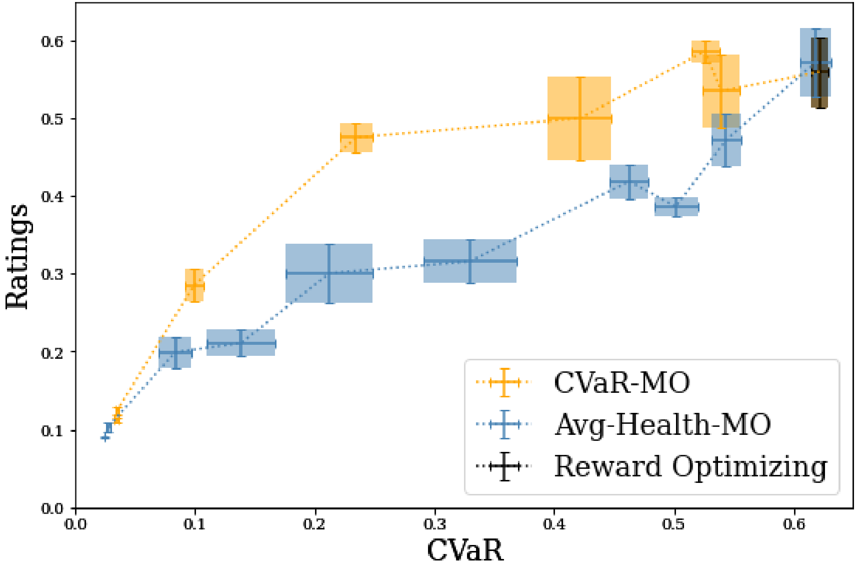
Idea: Some genres have a higher number of violent movies, and hence some users are vulnerable to more violent content under the reward optimal agent.

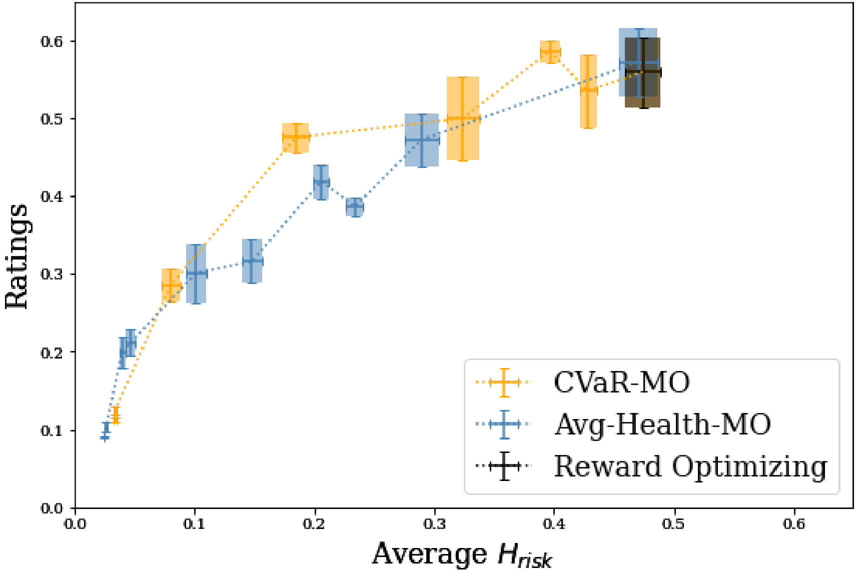
Distribution of $\mathrm{H}_{\mathrm{risk}}$ for different types of users

under the Reward Optimizing agent



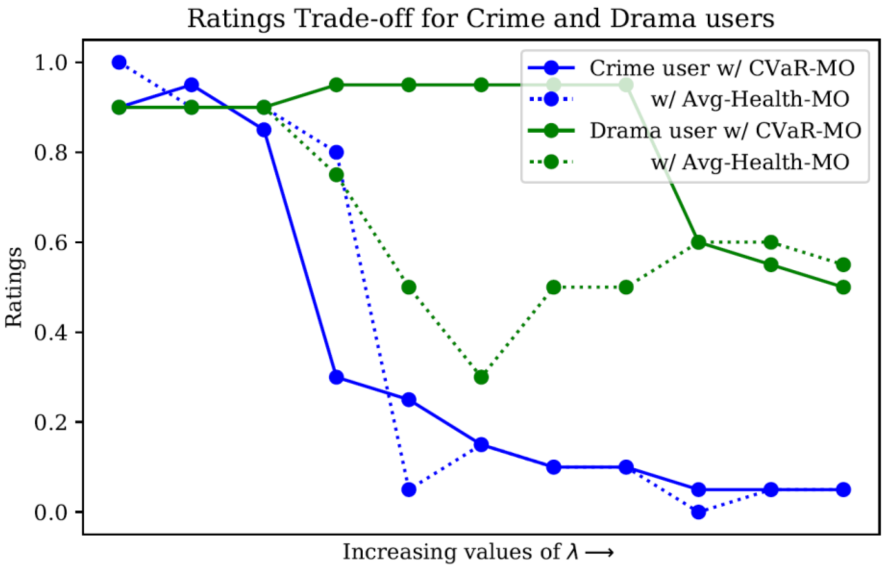
What is the trade-off between Health and Reward?

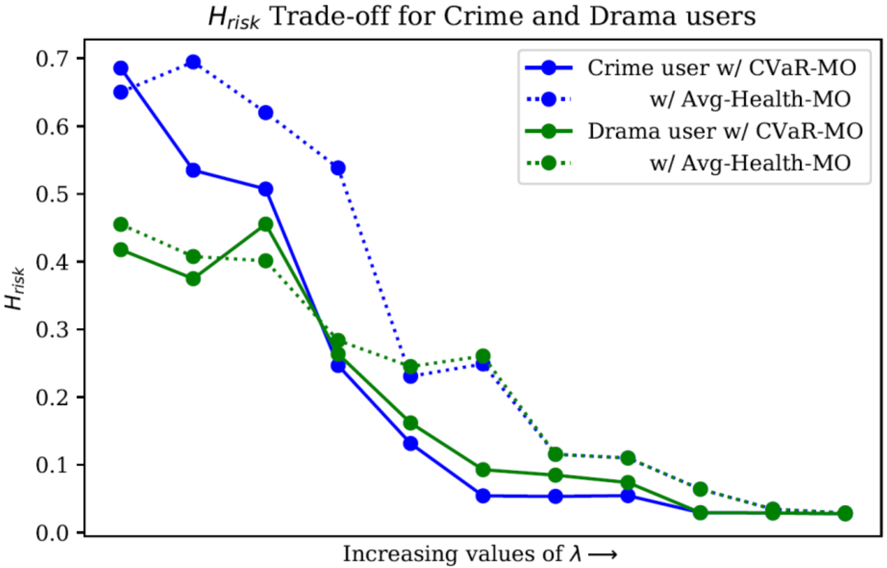




CVaR-Mo has a significantly better trade-off in terms of CVaR i.e. H_{risk} for the worst-case users.

How does the trade-off compare for different types of users?





Since Crime movies have a higher $\mathrm{H}_{\mathrm{risk}}$ than Drama movies, the CVaR-MO agent only trades off user ratings for the *Crime users* and not the *Drama users* (lower $\mathrm{H}_{\mathrm{risk}}$).

Looking forward

- Why optimize for the worst-case users?
    - Users experiencing higher risk might correlate with a certain population demographic.
    - Only optimizing the mean of the risk distribution is agnostic to what happens at the tail.
- How to characterize safety in a Recommender system?
    - Computing the aggregate health risk over each user’s trajectory.
    - Characterizing the tail of health risk distribution.
- Challenge: Identifying such phenomena in practical recommender systems in practice and performing simulations for research.

Thanks!

ashudeep@cs.cornell.edu

RecSys FAccTRec 2020

Building Healthy Recommendation Sequences for Everyone: A Safe Reinforcement Learning Approach

Ashudeep Singh, Yoni Halpern, Nithum Thain, Konstantina Christakopoulou, Ed H. Chi, Jilin Chen, and Alex Beutel

Cornell University

Google AI