NGRV NIGH GOLDENBERG RASO & VAUGHN

FAQ for MTR

Commented [1]: @dkolta@uber.com @andrew.hasbun@uber.com @anapaulapt@uber.com @dsvirsky@uber.com @gorkem@uber.com @bmarchi@uber.com @lfernandez@uber.com @gferreira@uber.com @duvlad@uber.com @mingxuan@uber.com

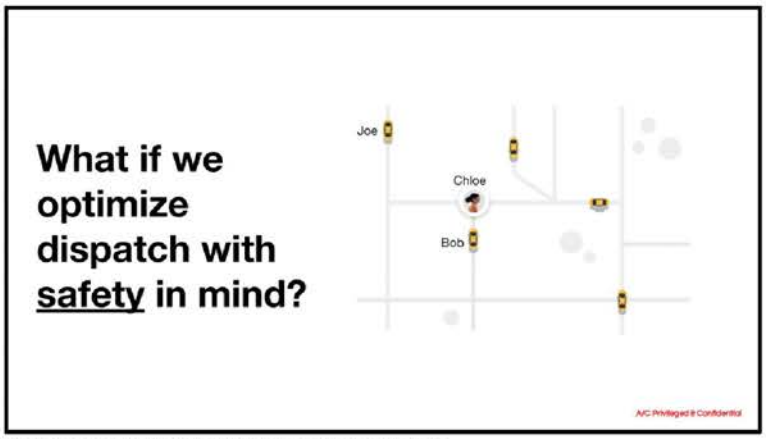
Hi Folks, We are crowdsourcing questions ahead of the MTR. Please take a look at the MTR deck (if you have not seen already) and then share any questions here that pop up from the deck or otherwise. Thanks!

Best,
Jenny

cc: @dharmin@uber.com

Resources: MTR Deck
Presentation: April 21

General questions



- **What do we currently optimize dispatch for?**
  Traditionally, we optimize for ETA and GB (long trips?)

- **We have been at it for a while. Why has it taken so long? / Why does it take so long to run a safety experiment?**
  Long experiment times are typical for safety experiments that try to measure incident rates for rare events. For S-RAD this gets even more pronounced since:
  - the product is aiming to reduce a subset of all safety incidents, making the target an even rarer event
  - the nature of the soft intervention brings to necessity to reduce the detectable effect size to levels as small as 5%

  Fo S-RAD:
  - We needed to conduct a bunch of prudence tests that give us confidence that SRAD is not breaking the marketplace, it is fair to various rider and driver cohorts, and that we could use gender inferences.
  - We needed to prove to ourselves that we are not simply shifting risk from one use to another, but actually making a safety impact. This required long experimentation, but our XP infrastructure was not geared to support our requirements; so we had to build a new infrastructure and then run the test.
  - Covid lockdowns and related trip volatility put the program on ice for a bit.

Trial Exhibit No.
03867

- **If it takes so long, why do we need to reach statistical significance? Shouldn't offline results or directional trends be sufficient? Why do we need to prove safety impact?**

  We have reasonable confidence that the model predicts risk
  We are testing whether the intervention works.
  Primary concern - Are we really reducing incidents or merely shifting risk from one driver to another or from one rider to another?

- **How did we determine different subcategories?**
  What we've learned from working with NSVRC, it is essential to make our taxonomy MECE, behaviorally specific, and easily interpreted by agents. That leads us to the most concise number of sub-categories as possible.

- **Why are some issues such as Sexual Misconduct and Verbal altercations considered non-serious issues? Shouldn't we have zero tolerance policy for Sexual misconduct?**
  All safety issues are unacceptable on our platform. However, what we've learned from working with the advocate groups, is that some behavior may be improved via training, while others require a more harsh policy.

- **What sort of sexual misconduct training resources do we provide/require for drivers and riders?**

Currently, we have sexual assault and sexual misconduct education for drivers in 51 countries and it's mandatory in about 19 countries. The materials were created by a third party expert, RAINN and other local NGOs in certain markets. Our safety education website, which I've listed next to the question, shows a preview of the lesson and a map on current coverage. It's linked to the source of truth which provide links to all the education content if there is a specific market you're interested in.

https://sites.google.com/uber.com/globalsafetyeducation/personal-safety-education

- **What is the difference in actions we take when we receive a non-serious versus serious versus critical sasm report?**

  For serious/critical incidents, we deactivate.
  For non serious incidents, we track the history and have a strikes system.
  We have education training, in some markets even mandatory.

Commented [2]: @alirorab@uber.com - can you please provide a tl;dr answer for MTR audience?

- **What are the 5 most severe sexual assault categories?**
  1. Non-Consensual Sexual Penetration (NCSP)
  2. Attempted NCSP
  3. Non-Consensual Kissing - Sexual Body Part
  4. Non-Consensual Touching - Sexual Body Part
  5. Kissing of a Non-Sexual Body Part

- **What are the top 3 serious sexual assault categories by volume?**

  Non-Consensual Touching - Non-Sexual Body Part
  Self Touching/Indecent Exposure
  Non-Consensual Touching - Sexual Body Part

- **How often do sasm incidents occur? Critical? Serious? Non-serious?**
  1 in 142K in US
  1 in <> in Brazil

- Which countries have the most serious SASMs?
  Brazil + US ~ 43%

- Ratio of Non-Serious to Serious
  Brazil - 4.8
  US - 6.3

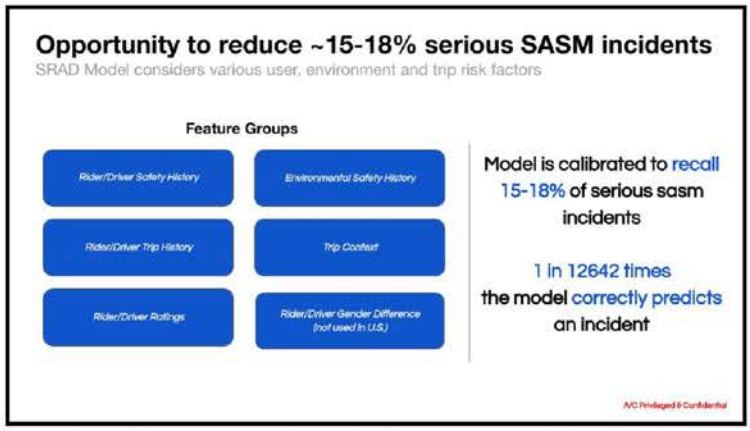
- **What is the difference between Sexual Misconduct & Severe Sexual Misconduct?**
  Severe SM: Masturbation / Indecent exposure

SM: Asking personal questions

- **Why did we decide to focus on serious versus all sasms?**
  Impact on the victim is much more significant
  The volume of non-serious is many fold higher than serious sasm
  Non-serious incidents could crowd out serious incidents in terms of learning for the model

- **Why did we decide not to focus on all IPCs?**
  We have other products such as SDM, Rider verification and DACT policies that help address non-sexual interpersonal conflicts.

  SRAD is complicated product; so we wanted to focus the team and model on Sexual assault problem first. Once we have tackled this, we intend to expand the footprint to all IPCs.



- **How do you decide the trigger rate?**
  We work with individual regional teams to determine at a trigger rate strikes a good balance between safety and marketplace metrics. Along these lines, we have conducted a number of experiments at various trigger rates to determine the impact on marketplace metrics such as p99 ETA, C/R, unfulfilled, cancellation rates, etc.

- **Is the threshold different for each city?**
  Yes, for city and time of day.

- **How do we calibrate the thresholds?**
  Thresholds are determined based on trigger rates agreed up with regional teams.
  We actively monitor the SRAD actions and update when we see actions outside the guardrails
  +/- 15% in country, and +/-25% in XL of the trigger rate

- **Is 15-18% recall the ceiling for the model?**
  No. If we increase % of trips affected by SRAD, we can increase the recall for the model. However, it also means that we will see a higher impact on marketplace metrics.

  Modeling improvements and new features being included may help improve the recall levels.

  We can think of model structure (consider sequential events like the ATO model) , *examples*: repeated request behavior, is the request only at specific times of day/specific locations

- **Have we seen patterns in SRAD trips flagged?**
  - Guest Rides
  - Intoxicated riders
  - Bar trips
  - Yet to be explored further

- **Why are we taking rider & driver features together? Why don't we assess their risks separately?**
  - SRAD is another way of looking at risk at the trip level
  - We believe fundamentally that the risk is situational - there is not something intrinsic of a rider or driver
  - Example: SDM looks at risk of rider (for theft, not SASM). We have DACT (look at riders or drivers based on safety incident history)
  - Dact models in H2 look at user risk.

- **What do we mean by rider/driver incident history?**
  Rate of safety incidents reported against a user
  Rate of sa/sm incidents reported against a user
  Count of incidents

- **What do we mean by rider/driver trip history?**
  Completion rates
  Cancellation rates
  Tenure of the user in terms of trips

- **What do we mean by ratings history?**
  Rate of 1 or 2 star ratings received by the user
  Rate of 1 or 2 star ratings given by the user

- **Do we use the user's background check history?**
  Not at this point; but we are continuously evolving the feature list to improve model performance. Some examples include - guest rides, fraud signals, etc.

- **If a user's incident history is bad, don't we ban them?**
  Our safety ops team has put together safety standard & policies that determine actions we take based on the severity, type and pattern of incidents. E.g. A report of non-consensual sexual pentration results in straight deactivation. But a pattern of non-serious sa/sm incidents take a couple of strikes to deactivate a user. Rog, do you want to add anything?

- **Which features have a higher influence on the predictions?**
  Every trip has a different relative feature importance.
  On the whole,
  - Driver incident rates
  - Client late night request rate
  - Client bliss ticket rate
  - Trip time
  - Trip city time of day historic SA rate
  - *Note: restaurant & bar count may not be in the Brazil model; these are in the US model*

Commented [3]: @bmarchi@uber.com @gorkem@uber.com

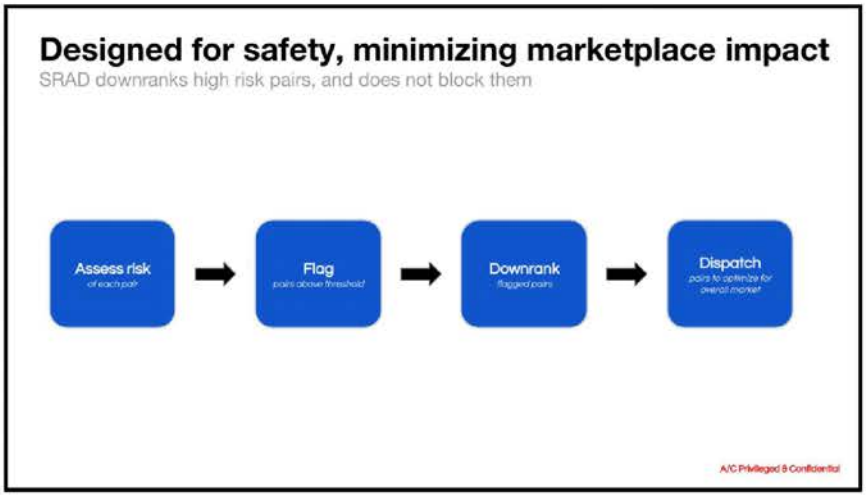
- **Why don't we use gender inference in the U.S.?**
  - In conversations with civil rights advocates in the US, the explicit use of gender raised concerns and legal risks. We removed gender to show that we were responsive to their point of view, but also because the model performed effectively without it. It's also important to note that removing gender doesn't remove the need to be vigilant for bias. A machine learning model can 'guess' gender if it wants to, which is why we continue to test for gender bias.
- **What is the problem with using gender inferences in the model?**
  - There were policy concerns and one legal concerns with using gender inferences in the model. First, in some states, explicitly using gender raised the scrutiny that the model would receive for disparate impact. Second, advocates felt it was disrespectful to guess someone's gender and use it in a model without getting consent proactively. As one put it – why not just ask people first?
- **What concerns did advocacy groups have?**
  - The main concern is simple discomfort with a strange technology – what right does Uber have to intervene in people's lives in a way that dictated who might or might not get sexually assaulted? And why do this in the background, without giving people a chance to weigh in?
  - Another main concern is racial discrimination. The model should not discriminate against any groups.
  - Secondary concerns were that users should be able to opt-out of gender and racial inference.
  - These variables would be securely stored and never used for marketing purposes.
  - Uber should be transparent about the use of the model and the PII used in it. Recommendations were to include it on the Privacy Notice, What Moves Us, and Download Your Data.
  - There should be external and independent audit of safety and fairness.
- **Which advocacy groups did we talk to?**
  - ReNika Moore, **ACLU Racial Justice Project**
  - Aaron Rieke, **Upturn and FTC**
  - Dariely Rodriguez, **Lawyer's Committee**
  - Gaylynn Burroughs, **Leadership Conference on Civil and Human Rights**
  - Ryan Budish, **Harvard Berkman Klein**
  - Diana Mancera, **Jane Doe** (Massachusetts Coalition Against SA and DV)
  - Sophie Hilgard, **Harvard Data Science Initiative**
  - Chelsea Barabas, **MIT**
  - Vincent Southerland, **Center for Race Inequality and the Law (NYU)**

- Karen Levy, **Cornell University**
- Roel Dobbe, **AI Now Institute**
- Amy Barasch, **Her Justice**
- John Verdi, **Future of Privacy Forum**
- Chad Sniffen, **National Alliance to End Sexual Violence**
- Erica Olsen, **National Network to End Domestic Violence (Safety Net Initiative)**
- Spandana Singh, **New America - Open Technology Institute**
- Liz Woolery, **Center for Democracy and Technology**

- **How is this model different from SDM?**
  SDM assesses a risk of a Rider committing a Theft or Robbery on the platform. When SDM risk is higher than the threshold, it blocks rider's access to the trip and we offer a set of rider verification challenges such as DocScan to the rider. If the rider passes the challenge, we grant them access.

  SRAD on the other hand assess the risk of a rider-driver pair. It does not block the dispatch, but downranks a pair that is deemed risky.

- **Why is the precision so low? How do we plan to improve the precision?**
  This is something related to the inherent unpredictability of the safety incidents. The team's efforts will continue to improve the precision of the model gradually, however the low-precision nature of S-RAD is not expected to change for the foreseeable future.



- **Shouldn't we just block drivers?**
  Low precision model
  Despite the high risk, the likelihood of an sa/sm on the trip is still very low
  Blocking drivers seems like a blunt tool to prevent incidents
  SRAD is not making a prediction on the rider or driver risk; it is making a prediction on the situational risk (ie time of day and city you are in)

- **If a rider is risky, would all pairs be flagged? Should we block the request at that point?**
  In future, we do plan to look into persistently actioned riders or drivers. If we can do so with high precision, we could take a stronger action.

- **What happens when all plans are flagged?**
  We leave some matches open to be dispatched.
  Find the least risky plans.
  Historically there have been incidents in a particular area.
  If we block trips, we are preventing provision of potentially safe ride for folks in that area.

- **Isn't it a legal issue if we know a trip is risky (high-score trip), but we still let it through?**
  SRAD offers a weak signal.
  We do not know if a trip is for certain (if we did, we would block these trips).
  We have a signal and are intervening based on that.

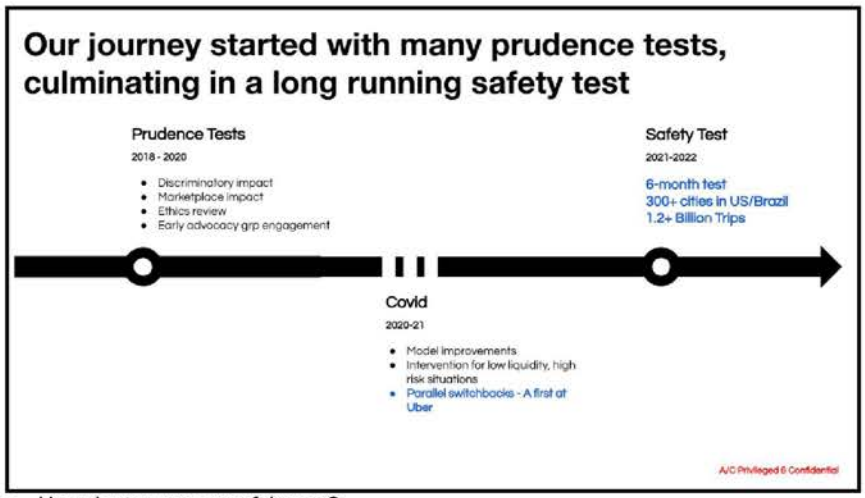
  Offline analysis shows us the potential.
  But because S-RAD is a soft intervention, we can't be sure that the offline recall translates into safety impact. For example, marketplace engine may ignore S-RAD's recommendations and still dispatch a high risk trip. Or there could be persistently actioned users who may ultimately escape the intervention
  Measuring the impact and understanding the behavior will help us improve the product and also help us achieve the desired balance between safety and marketplace impact.

- **How does SRAD work in conjunction with job boards?**
  Job boards shows trips that are reasonably available. If we have downranked pairs, then those jobs may not show up. This is something that we need to deep dive into.

- **What were the intervention improvements in low liquidity situations?**
  Our analysis indicated that most of the safety effect came from night time triggering. But we also had the opportunity to improve the model's effectiveness during day time so we increased the rate at which the model triggers during day time. From phase 5 analysis, we also found that almost 1 of 3 times S-RAD intervention was rendered ineffective because all S-RAD would flag all rider-driver plans generated for a given request. Because S-RAD used a soft-filter technique, the matching engine would ignore S-RAD's recommendations in such cases. So we introduced a Least-N risky intervention, in which S-RAD would not flag the least 25% of plans.

Commented [4]: @lfernandez@uber.com @gferreira@uber.com - do we know the answer to this?
Assigned to lfernandez@uber.com

Commented [5]: No, this is honestly the first time I hear about job boards. Do you have more context about it, so we can deep dive?



- How do you measure fairness?
    - For fairness testing, we make sure that S-RAD isn't leading to higher ETAs or lower earnings for different subgroups of users. We need to make sure, for example, that women don't have to wait longer for a trip on Saturday nights. That black drivers aren't seeing drops in earnings. That poorer neighborhoods don't see dips in service quality. And so on.
    - We work with civil rights and other advocacy groups to help us know where to look and what feels fair.

  Dan Svirsky

- **How do we know user ethnicity or which drivers identify as having non-binary gender?**
  Answer: On ethnicity inferences – we've done innovative work here. We partnered with civil rights NGOs and academics who focus on algorithmic bias to develop best practices for ethnicity inferences. The inference methods we use rely on users' names and locations. No single inference is perfect, so we've found that the best strategy is to use multiple inference approaches and see what they say directionally, taken together. That work is going to be public in June, and our hope is that other tech companies can use it, and that it's a chance for Uber to show leadership.

  On gender identity – some states allow drivers to list X as their driver's license gender, and some drivers call in and self-identify as transgender (we then give them specialized customer support). We can look at this population of drivers – about 10,000 in total – to see if SRAD is making things harder on them.

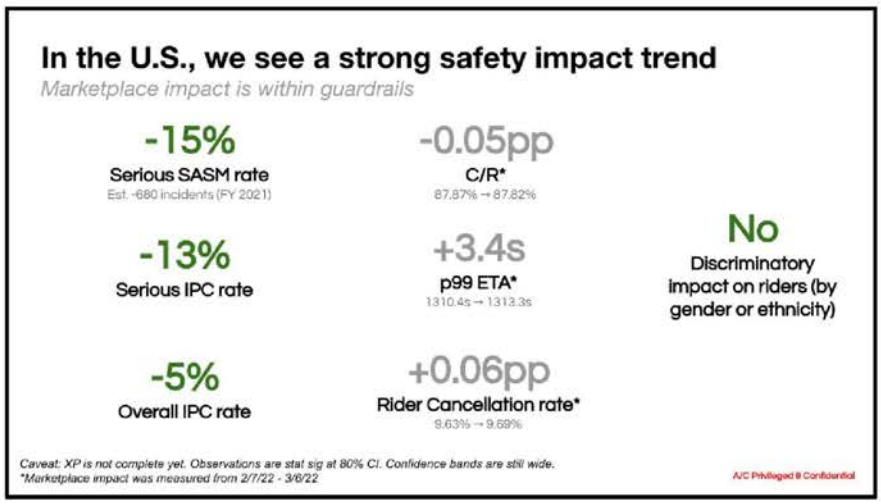
- **Do we look at race or ethnicity? Do we expect these features to change a lot from region to region when we expand?**
  Yes – we look at race and ethnicity in the US. We chose not to in Brazil, because we just don't feel comfortable with the amount of data we have. In the US, we can leverage

public work on ethnicity inferences. In Brazil, that doesn't exist to the same degree. That said, the Fairness Research team is working on solutions for international ethnicity testing and we hope to have that ready as SRAD and DACT continue to roll out.

- **Do users provide us with gender information? Do we have their consent to use the data?**
  In some cases, we do have user provided gender information. For example, we have driver gender information based on ID documents; but in cases where we don't have user provided gender information, we infer gender based on a user's name.

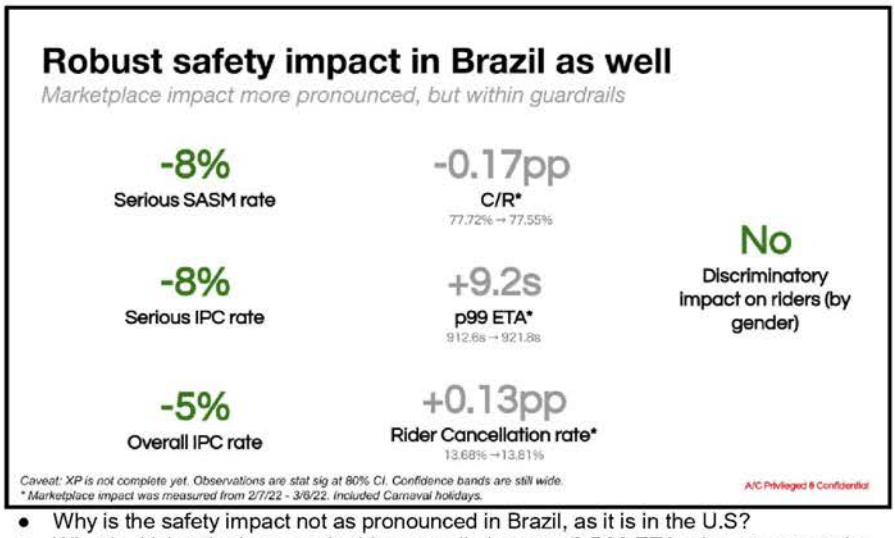
- **What is a parallel switchback?**
  It is a way to run more than one switchback experiment in the same city. A switchback is a form of experimentation that breaks treatment/control by time rather than users, used for experiments that could have network effects from treatment into control. Every 96 minutes, the whole city switches from treatment -> control and so on for 2 weeks - a careful design to guarantee balance between treatment and control and that the same slot in week 1 will have a different cohort than week 2. The problem is that with this design you can't run more than one switchback experiment in a given city, otherwise they'd interfere with each other. Uber has a platform for reserving cities for switchback experiments. Because S-RAD takes so long to reach stat sig, the pain of coordinating with other experiments made it impossible to rely on normal switchbacks. So we developed a parallel switchback, which essentially uses twice as big time slots, allowing a primary experiment to run along with S-RAD, eventually balancing them out

- **Is the parallel switchbacks infrastructure useful for non-safety?**
  - At this point, the benefit is limited to safety since it is probably the team that requires a long-running marketplace XP.

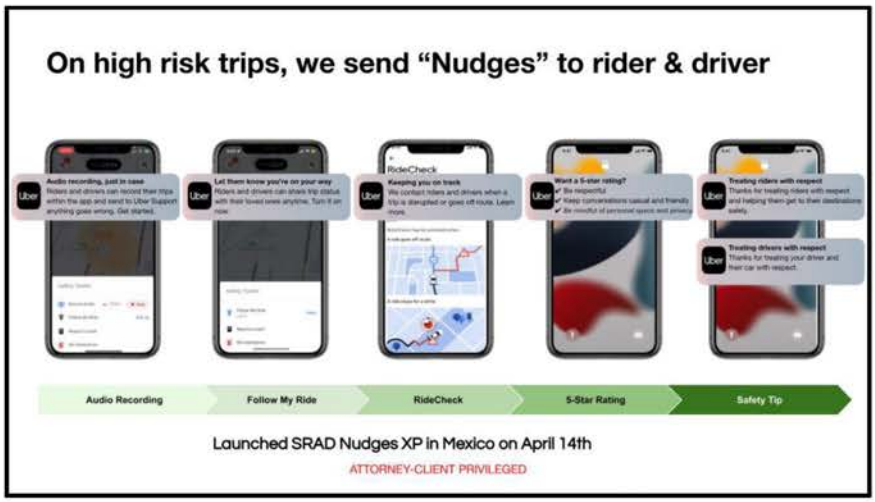


- What is p99 Request to Begin time?

-
- What is the relationship between flagging rate and marketplace metric?
- What inputs did you consider and remove?
- What marketplace metrics did you consider?
- **What knobs do you have?**
  - For each country, we can decide what % of trips we would end up downranking the best ETA plan.
    - In Brazil, the flagging rate is at 1.1%
    - In the US, the flagging rate is at 1.25%
- Why 80% confidence? Why not 95%?
- What is the impact on critical SASM rate?
- What is the reduction in top subcategories?
- Is there any impact on trip ratings?
- **What other marketplace metrics did you monitor?**
  - Unfulfilled rate?
  - Driver cancellation rate?
  - Night time?
  - Which cities are the most impacted?
- What does the 13% reduction in serious IPCs amount to?
- What does the 5% reduction in all IPCs amount to?
- Do you monitor spinner time?
- What are the guardrails for US?



- Why is the safety impact not as pronounced in Brazil, as it is in the U.S?
- What is driving the increase in rider cancellation rates? P99 ETAs do not seem to be increasing too much
- What are the guardrails for Brazil?
- Why is it taking so long to do a full rollout in US/Brazil?
- Why does it take so long to expand globally? Are we pushing hard enough?



- What is the long term strategy for nudges?
- What else can we do on risky trips?
- What metrics are you tracking?
- Should we consider blocking individuals at some point? Perhaps rule out access to night time / bar trips?
- Will there be an Explainability Tool to see how the rider/driver pair features are above/below the S-RAD trigger threshold?

Questions from Dara:

1. Does the model pick up features from riders or drivers more?
2. SASM rate for Brazil for the control group is lower than that of the US
    a. It doesnt match the reputation of Brazil
    b. Check that the core data is accurate
        i. The guess is that Brazil is much
        ii. Brazil is much worse for non-SASM
3. The algo takes into account cash vs non-cash trips
    a. Per Mingxuan: we have cash trips rate(client, geo) feature, while we don't take if the trip/request is cash/non-cash in real time into the model @dharmin
4. Will we have the capability to ensure that there is no gender and ethnicity impact?

a. Yes, we are building Holdout strategy to be able to set up infra to measure all the impacts
    i. Also bc models deteriorate
    ii. Holdout is controversial bc holdout means holding out and making trips unsafe for some people but it will be random
        1. We need to balance all the ethics
        2. Holdouts are blocks of trips, spread across the entire population, in a way that does not conflict with other MP xps
            a. If youre not holding out any particular population, Dara is okay.
            b. Could we share that we are A/R?

5. Can we talk about this?
    a. We are working on a reactive post
    b. Matt says its tricky and controversial
        i. Wants to guard against misinterpretation and criticism
        ii. Can build it into the safety reports going forward