August 28, 2021
Leslie Berland
Re: Health Objectives #1 and #2, and partnership with TwS
To: Peiter Mudge Zatko [redacted]

Ahh Mudge this is amazing. You deserve this type of bench! And this will be an amazing turn-around story in her career. Congrats my friend 🙏

On Fri, Aug 27, 2021 at 9:45 PM Peiter "Mudge" Zatko [redacted] wrote:

Leslie,

Please don’t forward Katrina’s note below.

I wanted to share it with you because it makes me extremely happy.

She’s precisely what, and who, we need to lead an *operations* team. Katrina is providing precisely the knowledge and expertise hoped for.

Her gravitational pull, executional maturity, and leadership is critical for improving (fixing) <cough> several teams.

Two of the 5 teams within confidence (privacy is the other) are now on their paths to being world class. Early and rocky but demonstrably on their paths.

I’ll continue to pour fuel on Privacy and TwS and bring the next confidence teams onto their paths…

Thank you for supporting and encouraging bringing Katrina onboard. I did my own research and legwork on her and felt comfortable but some others had reservations. Your support helped.

V/r,

Mudge

PS - I wasn’t expecting to need an entire rebuild in my world when I joined. But, damnit, if that’s what’s needed we’re gonna do it. :)

---------- Forwarded message ---------

From: **Katrina Lane** [redacted]
Date: Fri, Aug 27, 2021 at 7:47 PM
Subject: Fwd: Health Objectives #1 and #2, and partnership with TwS
To: Peiter Mudge Zatko [redacted]

TLDR: CONFIDENTIAL update for you on Health OKR partnership. As noted when we talked this week, we had a pretty heavy lift over the last several weeks to get from a high level conversation on "appoint someone with authority to force TwS to do this", to "TwS you need to just commit to the goals as we know the work", to an actual plan that we can validate against resourcing and start to make progress.   The most recent commitment push is demonstrated in the bottom email from Oscar.

We would never share this level of drama of course in how things stood. But I wanted you to know in case you ever have to defend the fact we stopped debating commitment and said we need to get a real plan (not what they wanted to hear). It is still a very real risk none of this moves the needs enough to meet their OKR but that was true before too and this plan will for sure move it more. We just don't know enough to size the impact until we do the first deliverables.

I have reviewed this with Oscar and I think he is starting to see the difference, but regardless likes the plan and realizes we may need their help to execute (though he would like our request in triplicate by Monday :) but we have work to do to get there. And importantly, I think our team really sees the difference and now feels set up to have an impact (though a long way to go in doing the work). This was also developed by them working as a team and pulling together - it's a very promising step.

For more details: To get a sense of what the difference is and the progress you can compare his asks in that email (theoretically ok but not actionable and no internal TwS alignment on details as the details were not documented), to the detailed current plan. You can also look at the fact that in his email the words and the actual linked pilot are not even doing the same thing (objective 2 ask 4) to get a sense of the disconnect. We bear some of the responsibility (the dynamics between the pillar leads caused a lot of this, and Michael's approach was also a big issue).   Health does too as they pushed very hard on commitment with a statement of fact that the data is clear and we know what to do that was both wrong and distracted all.

Let me know if questions and again if you would rather I spare you some of the sausage.
K

---------- Forwarded message ---------
From: **Katrina Lane** [redacted]
Date: Fri, Aug 27, 2021 at 7:56 AM
Subject: Re: Health Objectives #1 and #2, and partnership with TwS
To: [redacted] [redacted]
Cc: [redacted] [redacted], [redacted] [redacted],
[redacted] [redacted]

Please see our detailed update below getting to the specific actions we will be taking in partnership with parts of Health. We believe these plans are the right step to drive the metrics and also are at a more detailed and specific level which should help us collectively better align on the actual planned work. We have time today so we can discuss.

The team is proceeding to confirm named resources (though work is well in progress on the pilot and work will proceed on the others), as well as assessing any additional help needed from Health, and determining the risk to other existing health roadmap work in TwS (on the assumption that these OKRs are the priority). We should have a first cut to you by end of next week.

We took your original email and added in blue the new detailed plan. Talk soon...
k

Health Objective #1: Unhealthy impressions

In mode detail:

1. Develop a baseline to help us understand how much impressions "latency" we have in our system (the number of impressions that violative content accrues between posted and actioned, in this context time is irrelevant since tweets accrue impressions at vastly different rates)
2. Develop and implement queue/workflow ranking capabilities: We need to sort and prioritize the way cases are reviewing in a way that helps us achieve this goal. This work is inflight by Health Tools.
3. [TwS Owns] Launch an experiment by Sept 10th to validate that combining the TOS + Virality model results in reduced unhealthy impressions.Date for assessment will be sent by next week.

4. [TwS Owns] Assess options to align staffing to demand for specific health queues to determine potential impact on unhealthy impressions reduction by shrinking the report to action age gap. Surface options and requirements by end September, and collectively agree with Health what should be implemented within the quarter.
5. [TwS Owns] Run an opportunity analysis between now and end Sept to determine potential impact on impressions reduction to be realized by deprioritizing cases beyond Virality by age. We will start by assessing the impression decay timeframe for the xx biggest current health FIFO queues
   a. Identify how much agent capacity is currently allocated to FIFO queues
   b. Develop impressions vs time data for the x biggest FIFO queue and establish "half life" threshold. Then calculate the potential improvement in unhealthy impressions.
   c. Based on results: assess what it would take to operationalize this prioritization to the agents and agree with Health what can be implemented in year. Ideas include possibly an experiment with a subset to analyze potential impact or implement auto-closure/deprioritization across all FIFO queues
   d. *Commitments above may have external dependencies from Health XFN that TwS will need support for or require trade off discussions related to lower priority efforts.*
6. [**Request for TwS**] In the spirit of determining ROI, I fully recognize that impressions are not the only thing that matters (e.g. non-consensual nudity is higher severity than spam) and would like your help in developing a framework that we can align with T&S on to determining ROI:
   a. Our proposal - Health to ensure sure that we are pulled into efforts to create toxicity or harm/impact type frameworks (we are aware of several including severity of impressions chart in Trust and Safety) by being included as consulted in the DACIN, so that we can be sure that this is operationalized in a way to achieve the expected outcomes.
   b. Second: we will flag throughout this process anywhere we are seeing challenges or direction to pull resources away from our generally agreed to be highest harm work (CSE, NCN, PPI) in order to achieve numeric goals on overall impressions as we believe this would be an unintended consequence. We would like to agree on the best forums to do this on accelerated timeframes.

These replace the original request

1. *[**Request for TwS**] Align agent staffing times with the times at which impressions are generated on the platform*
2. *[**Request for TwS**] Ensure we execute **a pilot** to prove out that we can prevent a large number of impressions, and once we do so, shift a large number of agents towards proactive review workflows*
3. *[**Request for TwS**] Incorporate impressions into the criteria for how we evaluate the ROI of manual review (instead of just action rate)*

Health Objective #2: Customer expectations with regards to health-related interactions

In mode detail:

1. Top of the funnel:
    a. Set expectations (in the reporting flow) with customers that there are certain types of reviews (e.g. spam, annoying) in which they shouldn't expect a response from us
    b. Decrease the number of incoming reports by redesigning the reporting flow. This project is already in flight.
2. Middle of the funnel:
    a. Automate (primarily in the form of auto-closing with comms) with a particular focus on the large categories of cases (safety and media)
3. Bottom of the funnel:
    a. Ensuring that we respond to cases within our stated 3 days

Requests from TwS:

Status added below - in italics

1. Continue TSAN support to help us understand the entire funnel, determine appropriate pacing towards hitting goals, and identify the largest opportunities. *TWS: there are no plans to redeploy resources away from current health work but we need to document the current activities and resources and ensure that between existing resources/Health funded analysts that we are able to sustain this level. TSAN is working to document.*
2. Ensure Engineering and Health Data Science has access and training related to operational data and cases, this has been a critical gap that will help us understand opportunities to improve the front-end product *TWS: Per discuss, we need health to be more specific about what this would entail and if different from current RTB we will need to discuss how to resource/what would drop*
3. ADDED by TWS: We reviewed all existing planned actions against the rest of the funnel and commit to supporting two key middle of funnel efforts where we can help best ensure success: to autoclose bystander safety core reports with high likelihood of non-violation (with customer comms) and providing c/x response for reviewed/non-violative reports in the MTT reactive main group (Partially tracked in TSVC-5 and PAR-580)

Hi Katrina / [redacted] / [redacted] -

I realize there are some questions around Health objectives #1 and #2, specific commitments TwS is making in order to hit those objectives.

In the interest of getting to absolute clarity and ensuring we can promptly shift to implementation/execution, wanted to drop this email distilling the work needed down to the basics, for the full context, you can continue to refer to the H2’21 Health strategy.

One of my concerns is that once we factor in planning time, holidays, and any potential disruptions -- we have ~12 effective weeks left in the year to hit these goals (which is not a lot of time given the work ahead of us).

Happy to discuss any parts of this that are unclear, you disagree with, and/or require further refinement. Thank you 🙏🏾

Health Objective #1: Unhealthy impressions

We are reviewing and actioning the majority of violative content way too late into its lifecycle. Based on what we know about how impressions accrue, we have 2-4 hours before content accrues 50%+ of its predicted lifetime impressions, in the majority of cases we action content after it has already accrued the impressions. There are a number of drivers that contribute to this:

1. The way our workflows are ranked (mostly FIFO) because we lack more sophisticated ranking capabilities
2. Even in a world where we significantly improve ranking capabilities and front-ends - which takes time -, the review load is still split across two review platforms (RTP and ServiceCloud) improvements on one side don't carry to the other side
3. The mismatch in review capacity between when impressions occur on the platform and when agents are staffed to review them (e.g. weekends)
4. The way we evaluate success for workflows (primarily based on action rate)

Strategy in a nutshell: Review and action violative content **before** it accrues impressions

In mode detail:

1. Develop a baseline to help us understand how much impressions “latency” we have in our system (the number of impressions that violative content accrues between posted and actioned, in this context time is irrelevant since tweets accrue impressions at vastly different rates)
2. Develop and implement queue/workflow ranking capabilities: We need to sort and prioritize the way cases are reviewing in a way that helps us achieve this goal. This work is inflight by Health Tools.
3. [**Request for TwS**] Align agent staffing times with the times at which impressions are generated on the platform

4. [**Request for TwS**] Ensure we execute **a pilot** to prove out that we can prevent a large number of impressions, and once we do so, shift a large number of agents towards proactive review workflows
5. [**Request for TwS**] Incorporate impressions into the criteria for how we evaluate the ROI of manual review (instead of just action rate)
6. [**Request for TwS**] In the spirit of determining ROI, I fully recognize that impressions are not the only thing that matters (e.g. non-consensual nudity is higher severity than spam) and would like your help in developing a framework that we can align with T&S on to determining ROI

Health Objective #2: Customer expectations with regards to health-related interactions

When customers report content they believe violates policies to Twitter, it's a key moment in their journey and forms lasting perceptions as to whether Twitter "cares" and is doing something to improve the health of the platform. There are a few drivers that contribute to our current situation:

1. We don't do a good job setting expectations up front as to whether we'll be getting back to users on their report/appeals and if so when they can expect to hear from us, which results in:
    a. An inherent pressure to review as many reports as we can
    b. Inevitably disappointing some customers that expected a response
2. We don't have a clear unified sense of what makes a report "valuable" (is it about customer expectations? Is it about impressions?)
3. We rely on silent auto-closure of cases (28% of cases are silently auto-closed), which continues to perpetuate the perception that Twitter doesn't care

Strategy in a nutshell: Decrease the number of reports that we need to review, so we can focus effort on "high value" reports, appeals, and other requests and get back to users within 3 days

In mode detail:

1. Top of the funnel:
    a. Set expectations (in the reporting flow) with customers that there are certain types of reviews (e.g. spam, annoying) in which they shouldn't expect a response from us
    b. Decrease the number of incoming reports by redesigning the reporting flow. This project is already in flight.
2. Middle of the funnel:
    a. Automate (primarily in the form of auto-closing with comms) with a particular focus on the large categories of cases (safety and media)
3. Bottom of the funnel:
    a. Ensuring that we respond to cases within our stated 3 days

Requests from TwS:

1. Continue TSAN support to help us understand the entire funnel, determine appropriate pacing towards hitting goals, and identify the largest opportunities
2. Ensure Engineering and Health Data Science has access and training related to operational data and cases, this has been a critical gap that will help us understand opportunities to improve the front-end product