

AMENDED Exhibit 1183

**PLAINTIFFS' OMNIBUS OPPOSITION TO
DEFENDANTS' MOTIONS FOR SUMMARY
JUDGMENT**

Case No.: 4:22-md-03047-YGR

MDL No. 3047

In Re: Social Media Adolescent Addiction/Personal Injury Products Liability Litigation

How should we default new teens into new interactions settings?: a survey of safety and value

By [Diego Castaneda](#) and [REDACTED] (Both authors contributed equally to this Note)

This note explores how we bring in “safety” and “value” into how we develop and execute on a safer account model for teens. Ultimately, we hope to protect teens from unwanted visibility and interactions and, in particular, reduce bullying and combat child safety violations. In this note, we surface existing and new data signals regarding “safety” and “value”, and give recommendations for how we should construct a safer account model for teens.

TL;DR

- **Both value and safety are important.** Understanding the relationship between providing user **safety** (protecting teens from unwanted interactions) and facilitating user **value** (healthy social engagement) is vital for our team to make sound decisions about how we default reachability settings to teens.
- **DMs are where most unwanted interactions happen, but they can also happen in mentions and tags since there are no controls.** Research strongly suggests that DM's are the most problematic interaction vector for teenagers on Instagram. Moreover, teens delete approximately 47% of the DM threads and requests they receive in the first 14 days on Instagram within four weeks, indicating that a large number of unwanted interactions involving new users occur through this channel. While people are not reporting, untagging, or deleting mentions and tags nearly as often (less than 1/30th as often), we know from research and anecdotal evidence that unwanted tags and mentions are a serious problem for celebrities, top media producers, and other top accounts. Furthermore, there is evidence that users are not aware of these settings.
- **Teens want meaningful interactions from both people they know and people they don't, but they don't want inappropriate or spammy interactions.** Defaulting teens to a restricted DM setting could reduce DMs associated with new follows by 60% or more, thereby preventing an important way that new users form positive connections with people they don't know, as well as with accounts of people they do know who have unidentifiable information from their profile. It's likely that this decision would significantly impact growth. Multiple signals from qualitative research with teens support this -- teens tell us that they are fearful of missing out if unconnected DMs are restricted.
- **Healthy teen development.** Research in the field of teen neurodevelopment shows [that social engagement forms an important pillar in learning](#) to form healthy relationships with both existing and new friends, including strangers. External literature suggests that fostering positive relationships may mitigate the development of bullying behavior in teens.

Recommendations

In weighing the balance between “safety” and “value” for new teen accounts, **the signals surfaced in this investigation suggest that we keep the current default interaction settings as-is.** The value of DM's in promoting wanted social interactions form key pieces of a healthy Instagram experience and social graph growth in teens. Although, unwanted tags and mentions are a less serious problem for an overwhelming majority of users, we should still be proactive in meeting user needs and mental models of safety. While we feel good about balancing towards value, we can take proactive steps to honor when users desire safety.

- **Upsell intermediate better interaction controls to existing and new teens.** [Recent analyses of Better Interaction controls](#) indicates that users are using non-default controls and that the most useful option (among ‘everyone’, ‘off’, and ‘people you follow’) for reducing unwanted interaction is the intermediate option, “people you follow.”
- **Targeted blocking.** [In a recent experiment, the Messenger team](#) reduced adult-minor interactions by a targeted blocking of adults to friend unconnected minors when the adult has recently received negative feedback from other minors.

Assessing Safety and Value considerations

While we are working towards a decision regarding defaulting all new teens into private accounts (impacting visibility and discoverability), we also need to decide how (and if) to default new teens into restricted reachability settings for DMs, Tags, and Mentions.

Value and safety are both important. Understanding the how we provide user **safety** (protecting teens from unwanted interactions) and also facilitate user **value** (healthy social engagement) is vital for our team to make sound decisions about how we default reachability settings to teens. We hypothesize that by **defaulting into a limited reachability**, we will create a **safer space**, where unwanted interactions are limited. **We also recognize an inherent hazard in this approach.** Teens value interactions via DM— in interviews, they describe how these exchanges present opportunities to connect and grow healthy social circles and interests.

How do Safety and Value signals inform our POV on defaulting new teens into restricted settings?

Insights which support defaulting teens into restricted settings: ‘safety upside signals’

Insight 1. Data shows that people receive unwanted interactions in DMs

- **7% of teens** report experiencing bullying on Instagram each week.
- According to NETS user survey, **40% of all bullying experiences are reported in Direct Messaging.**
- B&H prevalence in DM's is very high (10%) **according to the September prototype metric**
- Severe violations targeted to those under 18 are known as a **high prevalence problem in DMs**

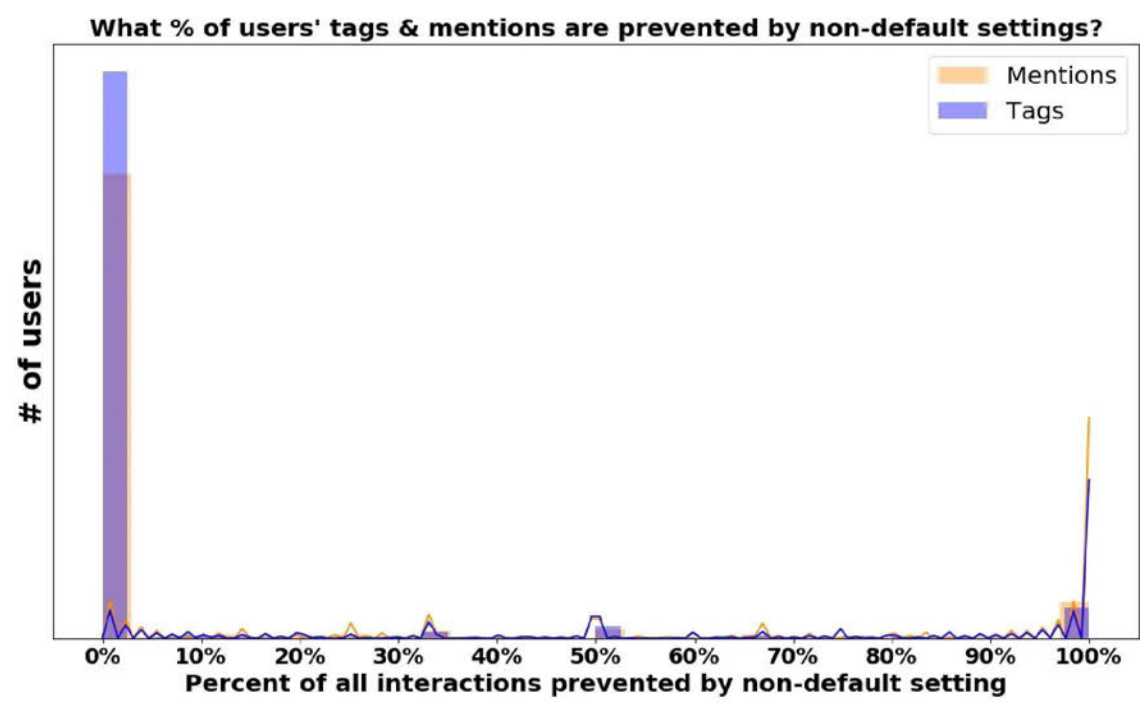
Insight 2. Starting new teens in a default restricted reachability setting could significantly reduce the volume of overall unwanted interactions in DMs, the primary vector by which people experience unwanted interactions.

- **New teen accounts (private or public) delete nearly half of all DM threads and requests over the first four weeks of their account.** 53% of these deleted DM threads and requests come from senders they don't follow. Of those *not* deleted, 55% come from accounts the recipient doesn't follow. For new users, threads they receive from unconnected accounts make up about half of both DM segments.
- **In general users signal unwanted interactions in DMs much more frequently than they do in tags or mentions.** About 1 in 3 newly created 1:1 DM threads are deleted by the recipient within a week. Less than 0.1% of mentions are reported and 2% of tags removed by their targets in the same time frame.
- **New users delete approximately 46% of all 1:1 DM threads and requests.** Ultimately, this means that ~24% of all 1:1 threads/requests are deleted and come from unconnected accounts and ~30% of all 1:1 threads/requests which are *not* deleted, come from unconnected accounts.
- **Strong association between blocking and deletion.** Approximately **6%** of all 1:1 DM threads and requests sent to new teen users result in the recipient blocking the sender within four weeks. However, about **13%** of DM requests deleted by the recipient result in the recipient/new user blocking the sender — *if a DM request/resulting thread is deleted, the probability of the recipient blocking the sender more than doubles.* In fact, more than 99% of all blocks involving a pair of users (new teen and sender) who had some DM interaction in the new user's first few weeks are associated with a deletion. Deletion in DMs is a strong signal of unwanted-ness for new users.

Insight 3. A test of new settings shows that restricted interactions settings can shield people from unwanted interactions.

Two weeks ago, the IG Bullying team began public testing of tags and mentions reachability settings ('everyone', 'people you follow', 'no one') in Australia and New Zealand. Though adoption is low, the test has provided useful information about who would use these settings and how they would provide value:

- **Users opting into these settings are encountering a disproportionate amount of integrity problems and that these settings present value for them:** 7% of users opting into these settings have submitted a report in the last four months, while only ~0.02% of DAP submit a report each day.
- **The people creating invalid mentions and tags are disproportionately bad actors relative to the general user population:** approximately 18% of invalid mentioners and taggers created at least one violating media in the last year. About 1% had at least one bullying violation in the same period.
- **Intermediate reachability settings provide value for some people:** more than 70% of users who made a settings change settled on a non-default option as their final choice. *About half of all settings users settle on 'people you follow' in the end, suggesting that this novel intermediate option is the most useful of the three.* It's clear that novel reachability settings provide value for some people. More than half of all invalid mentions and tags are attempted by actors completely unconnected to their target: *This suggests that restricted interactions settings are doing what they're supposed to do*, by preventing tags and mentions from random people unconnected to their would-be recipients.
- **Assuming that most of these preempted tags and mentions are unwanted, new interactions settings defaults could greatly benefit a small minority of users.** Approximately 85-88% of users opting into non-default mentions settings at least once in the first week of the test prevented **zero** mentions or tags using the settings during the week of testing. However, for ~8-10% of these users, the choice of non-default settings preempted 50+% of all tags or mentions they would have received during the period March 12- March 18. *Further analysis could help us better understand the people greatly impacted by choosing non-default interactions settings.*

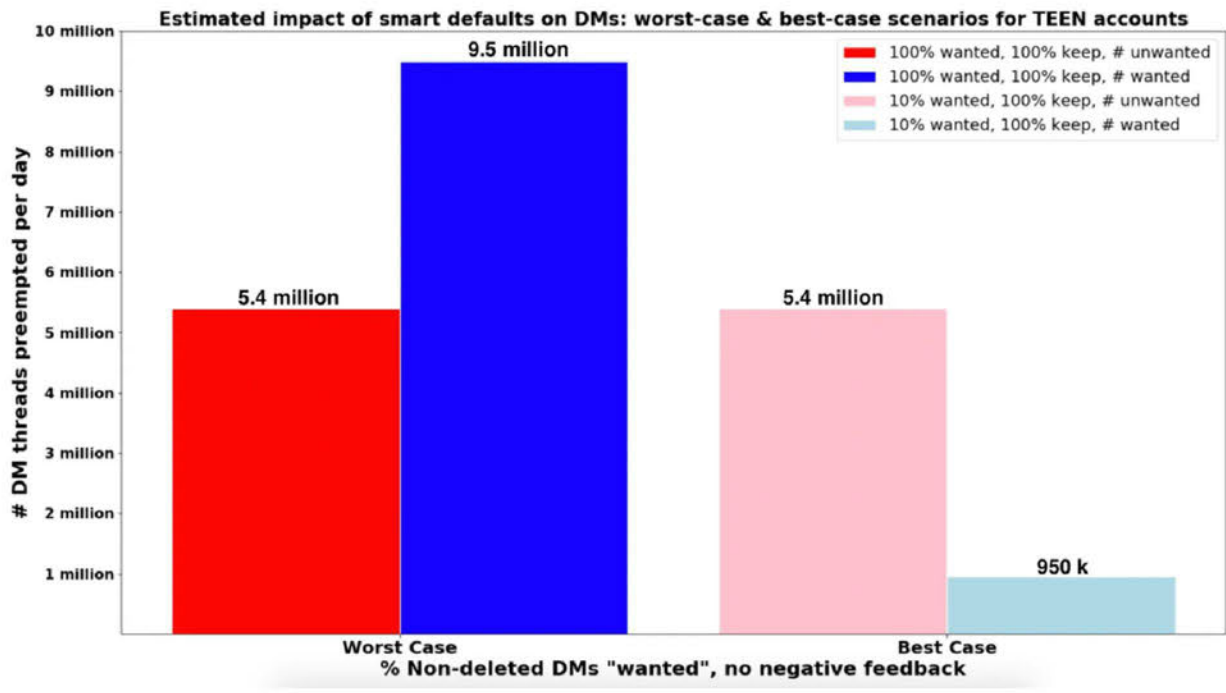


New interaction settings could greatly benefit a small minority of users

Insights which support NOT defaulting teens into restricted settings: ‘value downside signals’

Insight 1. Data projections show a strong potential for loss of valuable interactions in DMs.

For existing teen users, earlier analysis produced the following two scenarios of DM impact, assuming that we were to default all of these accounts to be reachable in DMs by only people they follow and that 100% of teens keep the change. The ‘worst case’ corresponds to the assumption that 100% of non-deleted DMs are ‘wanted’, while the ‘best case’ assumes that only 10% of these are actually wanted.



Analysis showing large potential loss of wanted DM's

Insight 2. Teens (and others) value connections gained and maintained through DMs.

- **Connecting with new friends via DM.** "A girl that I knew from school messaged me on Instagram. I ignored it at first because I didn't know who it was but then I figured out who it was. If I limited who could DM me then I would've never seen this."
 - Moreover, data shows that approximately 50% of 1:1 DM requests sent to new users in their first fourteen days on Instagram lead to the sender following the recipient at the end of the first 4 weeks. DMs from unconnected accounts are an important way that new users make connections.
- **Reconnecting with an old friend via DM's.** "I get a lot of spam DMs, but I can't just ignore them....Just last week someone from my middle school found me on Instagram and DMed me. I wouldn't want to miss things like this."
- **Reconnecting via tags.** "I don't want random accounts to tag me, but at the same time it could be a friend or someone at school that I just haven't connect with on Instagram yet"

"A girl that I knew from school messaged me on Instagram. I ignored it at first because I didn't know who it was but then I figured out who it was. If I limited who could DM me then I would've never seen this."

Insight 3. People will be confused if we change their interactions settings.

- **Consistency, Predictability and Control are super important for all users.** Bucket of users (e.g., everyone, people I follow, no one) should be the same for all interactions.
- **Meet expectations with respect to private or public setting.** People want [tag and mention] sub-controls to be defaulted to "people they know" for private accounts and "everyone" for public accounts.

"What if I'm reassessing the purpose for my account? This would include rethinking what all interaction look like. For example, I've decided I want to go public, maybe this means I want to have more interactions, so opening [tags and mentions] to "everyone" makes sense and is how I can get the most out of this."

- **People are unlikely to change their interaction settings on their own.** In the recent [Better Interactions test](#), only 4% of views of tags and mentions entry points result the user clicking through to view the new settings options. Of these, about 15-20% result in a settings change — ultimately, only about 0.6-0.8% of users encountering these entry points actually make a change.

Summary and Recommendations

What are the safety upsides for defaulting new teens into restricted reachability settings?	What are the value downsides for defaulting new teens into restricted reachability settings?
Data shows that people receive unwanted interactions in DMs	Data projections show a strong potential for loss of likely valuable interactions in DMs.
Starting new teens in a default restricted reachability setting could significantly reduce the volume of overall unwanted interactions in DMs.	Teens (and others) value connections gained and maintained through DMs.
Users opting into non-default interactions settings are disproportionately affected by integrity problems and a small minority (~1 in 10) significantly benefit from them.	We know that people value consistency and predictability in privacy related settings. Changing settings is likely to confuse and break their expectations.

In weighing the balance between “safety” and “value” for new teen accounts, **the signals surfaced through this investigation suggest that we should not change the current default interactions settings.** The value of DM’s in promoting wanted social interactions form key pieces of a healthy Instagram experience and social graph growth in teens. Although, unwanted tags and mentions are a less serious problem for an overwhelming majority of users, we should still be proactive in meeting user needs and mental models of safety.

While we feel good about balancing towards value, we can take proactive steps to honor when users desire safety.

- **Upsell intermediate better interaction controls to existing and new teens.** [Recent analyses of Better Interaction controls](#) indicates that users are using non-default controls and that the most useful option for reducing unwanted interaction is the intermediate option, “people you follow.”.
- **Targeted blocking.** [In a recent experiment, the Messenger team](#) reduced adult-minor interactions by a targeted blocking of adults to friend unconnected minors when the adult has recently received negative feedback from other minors.

Next Steps

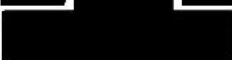
Additional questions to answer with data:

- *How should we handle other teen accounts?*
- *How urgent is it for us to pursue entry points and contextual education of our interaction settings?*
- *How can we better understand the people who would most benefit from non-default interactions settings?*

Conduct Research

- **Conduct user research to understand public private user journey.** Having a private & public journey could help highlight [insights](#) around how both of these account types (private & public) need controls around interaction settings and they depend more on the individual than the account type.

Acknowledgements

    
 [Shilpa Mody](#), 

References

- [Better Interactions: testing tags and mentions settings in Australia and New Zealand](#)
- [Reducing IIC victimization by restricting adult--> minor friending](#)
- [Survey on perceived reach + intensity of 18 negative experiences on Instagram](#)
- [Unified Prevalence Measurement for Bullying & Harassment - POC Update](#)

  15

14 Comments 4 Shares

 Like

 Comment

 Share



cc:  [Max Eulenstein](#) 

Like

Reply

1y

 5

#thanks **Diego Castaneda** and [REDACTED] for all your hard work on this note. It's a very difficult space and you've both done an extremely thorough job weighing the pros and cons.

Like Reply 1y



2

cc [REDACTED]

Like Reply 1y



5

Like Reply 1y

[REDACTED] **Diego Castaneda** based on your summary, it seems like we can measure whether we're reducing unwanted interactions by seeing if there is a reduction in deleted DMs, do you agree with that? Do we have confidence that reporting would also decrease if we are reducing unwanted interactions?

If we think about how we can measure "value" going forward, wondering if we can use these signals based off your note:

- back and forth messages between two parties
- messages that result in follows (do we have any signals that tell us what types of messages result in a follow?)
- follows, which we know is important for MAP

Any other ideas?

If we can align on how we're measuring whether we're decreasing unwanted interactions while maintaining value, we can better determine how to support these tradeoffs.

cc [REDACTED]

Like Reply 1y



Diego Castaneda

thanks for the comment [REDACTED] and the ideas around measuring value. To your questions

- **it seems like we can measure whether we're reducing unwanted interactions by seeing if there is a reduction in deleted DMs, do you agree with that?**
- **Do we have confidence that reporting would also decrease if we are reducing unwanted interactions?**
- Yes, based on Aaron's analysis DM deletion is a strong signal that the DM is an unwanted interactions. That's pretty clear. **"Deletion in DMs is a strong signal of unwanted-ness for new users"**
What is not as clear is the level of severity of each of those unwanted deleted DM's. its likely that most are just spammy and annoying and that people are willing to trade them off for the valuable DM that they are expecting or get serendipitously. that being said..we know that deletion of a DM is associated with some eventual blocking, indicating severity. Does blocking count as reported prevalence?

Your suggestions

- **"back and forth messages between two parties"**
"Moreover, data shows that approximately 50% of 1:1 DM requests sent to new users in their first fourteen days on Instagram lead to the sender following the recipient at the end of the first 4 weeks. DMs from unconnected accounts are an important way that new users make connections." (or course, we have to be careful with the small but very severe violations that can occur during these messages between adult and child.
- **"messages that result in follows (do we have any signals that tell us what types of messages result in a follow?)"**
we don't. thats why we are launching this research I think the research we are doing [with the new teenage user journey](#)(POC: [REDACTED] can give us signal on that.
- **"follows, which we know is important for MAP"**
Probably need more granularity or add it to other indicators of pro-social relationship building. ie do those follows build on each other?



Connect Quip account to see this preview

Like Reply 1y Edited



[REDACTED] **Diego Castaneda** awesome! Super excited to see the results of this. I may have missed it in the quip, but when do we expect to have these results?

Also, confirming that we'll ask these teens what they do after they receive an unwanted interaction? (such as delete or block?)

Like Reply 1y



View 1 more reply

FYI **Karina Newton** in case she hasn't seen this yet

Like Reply 1y



2



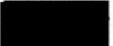
Karina Newton

On the second recommendation **Diego** and [REDACTED] the messenger integrity team is looking at porting over the same restrictions that FB has (unconnected adults can't message minors) however since we only have age for slightly more than half of users, we'll have to rely upon re-trained age classifiers later this half which is a) not ideal b) not a good IIC protection because groomers frequently build graphs that mirror teens. Which is all to say, there is a lot more targeted work that FB can do with respect to age and interaction to ensure safety than IG can and why there is more pressure on IG to have safer default states for new users (which is a rare case where we have age of 100% of those IG users). I have a hard time looking at numbers showing new teens delete half their their threads (I would also be interested in knowing if there's a gender breakdown on that -- based on feedback we've gotten, I would be surprised if teen girls aren't deleting at stat sig higher rate than teen boys), and thinking up-selling settings is an adequate response to the status quo. cc [REDACTED] **Vaishnavi J**

Like Reply 1y

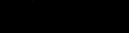


3

Diego Castaneda 
Karina thanks for the comment and mindful insight about the age challenges we deal with. Cc'ing  regarding DM and gender differences.

Like Reply 1y



 
cc.  ,    note "Assessing Safety and Value considerations" where we weigh benefits.

Like Reply 7w

 
Diego Castaneda This is awesome! Thanks for sharing.

Like Reply 7w

 
cc  as we were discussing (teen) unwanted messages this week

Like Reply 6w Edited



 Write a comment...

Press Enter to post.

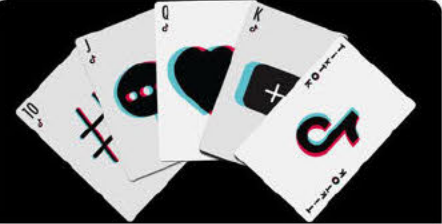


Recommended Reading



Refocused Instagram Org Structure

 Adam Mosseri



Reflection on building products at TikTok



Be Different

 Naomi Gleit