Why Bans Fail: Tipping Points and Australia's Social Media Ban*

Leonardo Bursztyn[†] **Angela Duckworth**[‡] **Rafael Jiménez-Durán**[§]
Aaron Leonard[¶] **Filip Milojević**[‖] **Christopher Roth**[**] **Cass R. Sunstein**[††]

April 27, 2026

Abstract

In December 2025, Australia became the first country to ban youth under 16 years old from holding accounts on major social media platforms, a policy now under consideration in more than a dozen countries and in numerous states. Because social media use is inherently social, the effectiveness of a ban that is easy to circumvent may depend on whether compliance reaches a tipping point: a share of compliant peers high enough to make it optimal for individuals to comply themselves. We surveyed 746 Australian teenagers four months after the ban took effect and find that only about one in four 14–15-year-olds comply. The social environment around use has barely moved: most banned teens believe that their peers are still using banned platforms and cite social reasons for continuing use. Sustaining high compliance requires two ingredients: the share of compliers must be high enough and those who comply must find it preferable to continue complying. The current ban achieves neither. Teenagers report that they require roughly two-thirds of peers to stop using social media to stop themselves, far above the share currently complying. They also perceive compliers as less popular than non-compliers, so the more influential teens disproportionately stay on the platforms. Together, these patterns suggest that compliance is more likely to diminish than to rise. Sustaining higher compliance will likely require pairing the ban with instruments that act on social norms and individual incentives directly.

*We thank Simon Cordes, Ingar Haaland, Luca Henkel, Lukas Hensel, Matt Lowe, and Shakked Noy for very useful and constructive comments. The research described in this article was approved by the University of Chicago Social and Behavioral Sciences Institutional Review Board. Duckworth acknowledges financial support from the Walton Family Foundation. Bursztyn is founder and CEO of NOMO Technologies, Inc., a company that develops digital wellness tools, and holds equity in the company. Duckworth serves as a scientific advisor to NOMO Technologies, Inc. and holds equity in the company. NOMO had no role in the design, conduct, funding, or analysis of this research.

[†]University of Chicago and NBER, bursztyn@uchicago.edu.
[‡]University of Pennsylvania, duckworth@upenn.edu.
[§]Bocconi University, IGIER, CEPR, CESifo, and Stigler Center, rafael.jimenez@unibocconi.it.
[¶]University of Chicago, aaronleonard@uchicago.edu.
[‖]University of Chicago, milojevic@uchicago.edu.
[**]University of Cologne, MPI for Behavioral Economics, and CEPR, roth@wiso.uni-koeln.de.
[††]Harvard Law School, csunstei@law.harvard.edu.

1 Introduction

In December 2025, Australia became the first country to ban those under the age of 16 from holding accounts on major social media platforms. Comparable legislation has since been adopted, drafted, or proposed in more than a dozen other countries and in numerous states. Bans of this design respond to widespread concern that heavy use of social media harms adolescent mental health (Braghieri et al., 2022; Allcott et al., 2020; Haidt, 2024). An important question is whether such bans, which are easy to circumvent, achieve their intended goal. The value of using social media depends on who else is on the platform (Bursztyn et al., 2025b,d), meaning that at high peer use the individually optimal choice may be to continue using even when access is formally restricted. Whether a ban can sustain itself therefore depends on individuals' compliance thresholds—the share of peers each teen requires before complying themselves—and how those thresholds are distributed across the population (Schelling, 1971; Granovetter, 1978; Braghieri et al., 2026).

Unlike restrictions on tobacco or alcohol consumption, for which enforcement operates at a physical point of sale (DiNardo and Lemieux, 2001; Carpenter and Dobkin, 2011), social media platforms can be accessed from any device at any time, and the Australian law imposes no penalties on individual users. Three main channels could shift behavior under the ban: perceived sanctions to non-compliers, increased difficulty of access to social media, and a change in the peer environment that raises the social cost of non-compliance with the norm expressed by the law (Sunstein, 1996; Bénabou and Tirole, 2026).

To measure the ban's effects and the channels through which it operates, we first surveyed 507 Australian teenagers aged 14–18 several months after the ban in March and April 2026, together with a contemporaneous sample of US teenagers and an additional mechanism survey fielded on Australian teens in April 2026. We find that compliance is low: approximately 27% of banned 14–15-year-olds comply.[1] The peer environment that would consolidate this compliance has barely moved. Most banned teens believe their peers are still using banned platforms, most describe circumvention as easy, and most non-compliers point to social forces—friends still on the platforms and the fear of missing out—as the reason why they continue. Compliers find it harder to keep up with friends and report feeling more bored. Thus, the pull toward the platforms has not weakened, consistent with the strong network effects that shape substitution in this market (Bursztyn et al., 2025d). This raises a natural question: what share of peers complying do teens require before complying themselves and how does this compare to observed peer compliance?

A companion mechanism survey estimates that share directly. Respondents are asked about their beliefs regarding current peer compliance and then state the share of peers required for them to comply themselves. Across different framings of the threshold elicitation, the mean stated threshold always lies around three-quarters, well above observed compliance (27%). The thresholds teens report imply that compliance at the rate currently observed cannot sustain itself. Combining stated thresholds with current peer-compliance beliefs, the only level at which compliance would be self-sustaining—where the share of teens whose threshold is met equals the share complying—is around 18%, far below today's rate. Given that the current compliance level is substantially below

[1] In the US, the share of teens using social media remains high and fairly constant across ages.

two-thirds, compliance is more likely to erode than to rise going forward.

A potential explanation for this pattern has to do with the social composition of compliers vs. non-compliers (Bénabou and Tirole, 2006). Our survey evidence shows that teenagers perceive those who comply with the ban as less popular than those who do not. As a result, the teens with the greatest social influence are currently more likely to stay on social media, keeping the platforms the "cool" place to be. Note the contrast with cigarette smoking among young adults, in which connected groups quit together and continued smokers progressively became peripheral in their social networks (Christakis and Fowler, 2008). This contrast points to a more general observation: sustaining a high-compliance equilibrium requires two ingredients, not one. The share of compliers must be high enough and those who comply must find it preferable to keep complying. Whether the composition of compliers and non-compliers shifts under sustained enforcement, and whether norms eventually realign, will determine whether the policy can reach a high-compliance equilibrium.

Finally, we discuss alternative and complementary policy changes, such as caps on time usage (as opposed to a ban), school grade-based bans (as opposed to age-based bans), promotion of alternative forms of peer interaction, and direct individual incentives for behavioral change. Beyond these instruments, we also emphasize the importance of social norms. A ban is more likely to be effective if the social costs to non-compliance increase.

The paper proceeds as follows. Section 2 provides background information about the Australian ban. Section 3 provides evidence on the current partial compliance. Section 4 describes the potential mechanisms which may sustain this low level of compliance. Section 5 presents the evidence on individual thresholds and norms surrounding compliance. Section 6 concludes with a policy discussion.

2 Background: The Australian Social Media Ban

The Online Safety Amendment (Social Media Minimum Age) Act 2024, passed on 29 November 2024 and in force from 10 December 2025, sets a mandatory minimum age of 16 for holding an account on ten major social platforms: Facebook, Instagram, Snapchat, TikTok, Threads, X, YouTube, Reddit, Twitch, and Kick (Parliament of Australia, 2024; eSafety Commissioner, 2026). The banned platforms all enable social interaction between users and allow them to post content.

Messaging services such as WhatsApp and online games such as Roblox or Steam are excluded. Parental consent does not bypass these restrictions, which is a common workaround used in youth restrictions for other goods.

Two features of how the Act is enforced directly affect the private cost non compliers actually face. First, enforcement is fully through platforms: companies face civil penalties of up to $49.5 AUD million for failing to take "reasonable steps" to prevent under-16 accounts, while minors themselves face no legal sanction (Parliament of Australia, 2024). The detection methods platforms have used—facial age estimation, identity verification, behavioral inference from language and login patterns, and signals from peer networks (eSafety Commissioner, 2026)—are imperfect. Thus, many teens can continue using their existing accounts. Second, those who had accounts removed

can typically open another since the Act does not address device-level access, where circumvention is possible. VPNs, false birthdates at sign-up, and accounts borrowed from older siblings or parents all leave the user with platform access at low individual cost.

The Act therefore raises the cost of holding a *detectable* under-16 account but leaves the cost of *continued use*—to a teen willing to circumvent—close to where it was before. Whether the ban changes behavior at scale depends on how these costs compares to the value of staying on social media, which itself depends on what peers are doing. The remainder of the paper addresses these points.

3 Compliance with the Ban

This section presents descriptive evidence on the share of Australian teenagers subject to the ban who are complying four months after its implementation.

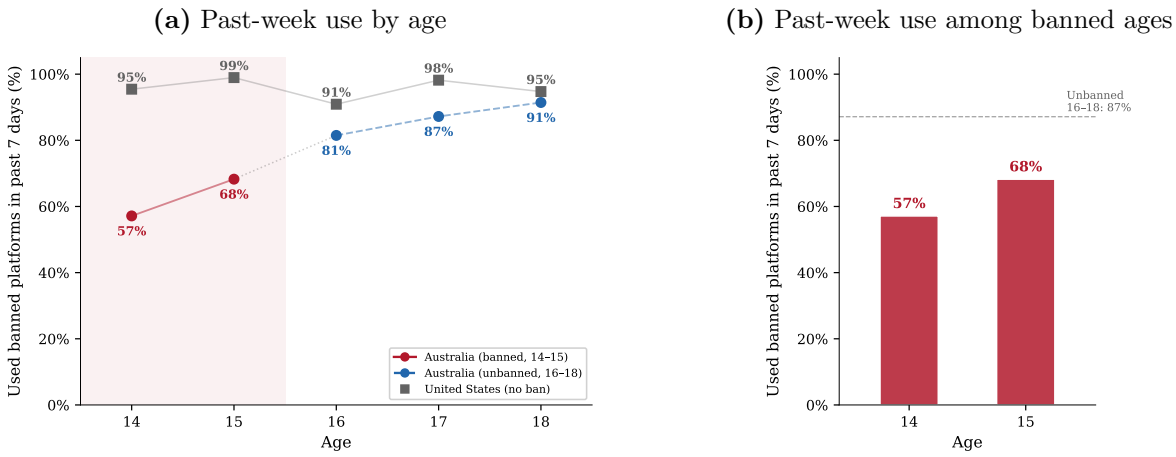
Sample We surveyed 507 Australian teenagers aged 14–18 between March 8 and April 6, 2026 through Youth Insight, a panel with broad coverage of Australian secondary school students. 45% of respondents are aged 14–15 and therefore subject to the ban; 48% are male; 92% are currently enrolled students. We also surveyed 300 US teenagers aged 14–18 through TeenVoice as a comparison group facing no comparable policy. Recruitment details and sample characteristics can be found in Appendix A2.

Survey design The instrument elicits respondents' awareness of the ban and the platforms it covers, their own recent social media use and beliefs about their peers' use, their reasons for continued use, the circumvention methods they use or have observed, their stated policy preferences over alternative regulatory tools, and self-reported wellbeing under the ban. These survey questions appear in Appendix A4.

Compliance rates We find that non-compliance is widespread. Among 14–15-year-olds subject to the ban, 63.8% report past-week use of a banned platform (Figure 1, Panel a). The corresponding rate is 87.1% among unbanned Australian 16–18-year-olds. Taking the unbanned Australian rate as the counterfactual, we estimate an aggregate compliance rate of roughly 27% (i.e. 63.8% of 87.1% teens do not comply). There is substantial heterogeneity by age: 57.1% of 14-year-olds versus 68.3% of 15-year-olds report using banned platforms in the last week (Figure 1, Panel b).

We consider two robustness exercises on this estimate. The first concerns the counterfactual in the denominator. Among unbanned Australians, use rises with age (81% at 16, 87% at 17, 91% at 18), so 87.1%—the 16–18 average—may overstate what 14–15-year-olds would have used absent the ban, in which case the implied compliance rate would be biased upward. Using US 14–15-year-olds (97.5%) as a same-age but cross-country benchmark gives a higher figure of 35%, though cross-country differences make this benchmark imperfect in its own way. The second robustness check concerns measurement of the numerator. Self-reports of own use may be subject to social-desirability bias if teens perceive—correctly or not—that admitting use implicates them in illegal

Figure 1: Social media use by age, Australia and United States



Notes: Figure 1 reports self-reported past-7-day use of any banned platform. Panel (a): use rate by single year of age. Australian 14–15-year-olds (red) are subject to the ban; 16–18-year-olds (blue) are not. US teenagers (gray) face no comparable policy. Use is lower at the banned ages in Australia but approximately flat across ages in the US. Panel (b): past-week use among banned 14- and 15-year-olds, on the same scale as Panel (a). The dashed reference line marks the unbanned Australian 16–18-year-old average (87.1%), the counterfactual against which the implied compliance rate of 27% in the body is computed.

behavior. We address this directly through a randomized-response module (Warner, 1965) that preserves individual plausible deniability while recovering group-level prevalence. The randomized-response estimate of past-week use is 64.5%, essentially identical to the direct 63.8%.[2] We read this as evidence that 63.8% is a valid measure of actual use, and as an indication that teens do not anticipate social costs to admitting continued use after the ban—consistent with the near-universal awareness, documented below, that the Act is enforced on platforms rather than on individual users.

Our headline figures are in line with three independent surveys fielded in the same period: the eSafety Commissioner's March 2026 compliance report finds 63–69% of under-16s retained accounts on Facebook, Instagram, Snapchat, and TikTok (eSafety Commissioner, 2026); the Molly Rose Foundation reports 61% of 12–15-year-olds still actively using platforms (Molly Rose Foundation and YouthInsight, 2026); and a YouGov survey of Australian parents finds most observed some behavioral change, with a substantial minority reporting migration to alternative platforms (YouGov, 2026).

4 Why is Compliance Low?

A ban on a highly social good can lower use through three channels: a personal sanction that raises the cost of being caught, access frictions that make social media use more difficult to achieve, and a social environment in which the attractiveness of non-compliance depends on what others do. We now discuss these channels one by one.

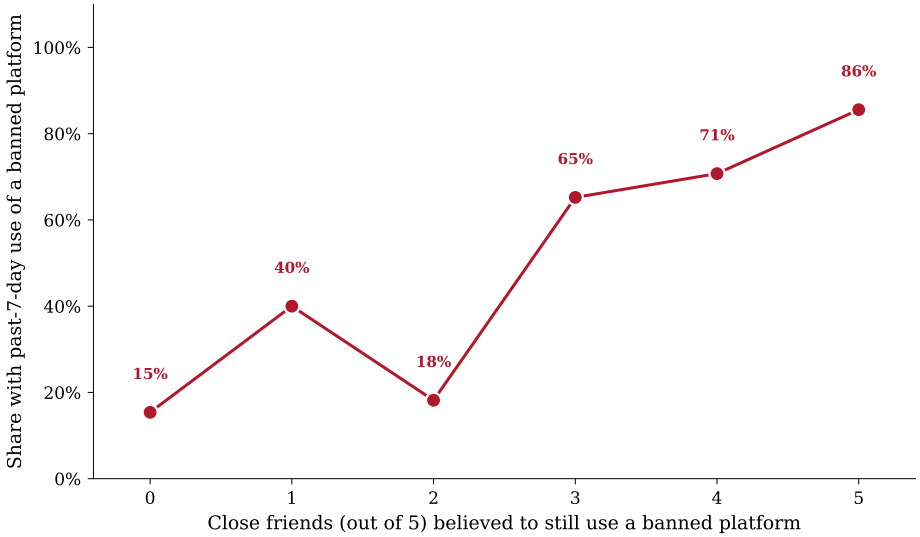
[2]The RRT estimate should be interpreted with caution given increasing evidence that respondents often misunderstand the randomization mechanism, which can bias recovered prevalence rates (Bursztyn et al., 2025c).

Perceived sanctions Personal consequences for non-compliance are not salient to teens. Only 22% of banned teens believe there are personal consequences for violating the ban; 47% believe the consequences apply to platforms alone—a perception aligned with how the Act is actually enforced. Awareness of the ban itself is near-universal: 86% know which platforms are banned. The lack of perceived personal cost is therefore not an information gap but an accurate reading of the enforcement architecture, in which platforms bear the legal risk and individual users do not.

Cost of access We find that 75% of banned teens consider circumventing the restrictions easy or very easy. In part, this is because platforms have not identified who are the under 16s: 64% of 14-15 year olds in our data have not had their social media account removed. As a result, for these users, the ban has minimal impact on the cost to continuing social media use. Further, for those impacted, the most common workarounds are entering a false birthdate at sign-up (44%), lying about age on verification prompts (57%), using a parent's or older sibling's account (42%), and routing traffic through a VPN (30%). These actions are no to low cost from a financial perspective, as even some VPN providers are free or offer free trials. Additionally, 25% of non-compliers report that a parent, older sibling, or other adult helped them sign up for a social media account after their personal account was deactivated. The cost of access has not meaningfully risen for a teen willing to use any of these workarounds.

The social environment Continued use is, on respondents' own account, driven primarily by social forces: among teens still using social media, 52% cite at least one social motive, with 42% reporting that their friends still use the platforms and 27% citing fear of missing out. We further test for the potential role of the social environment by splitting banned 14–15-year-olds at the median in the number of their five closest friends believed to still use banned platforms: 80.9% of those above the median report personal past-week use, against 36.5% below ($r = 0.50$, $p < 0.001$). Figure 2 shows that own use rises smoothly with the number of close friends believed to still use banned platforms across the entire gradient. The correlation is descriptive—peer effects, social learning, and homophily are each consistent with it—but its sign and magnitude suggest that own behavior tracks perceived close-peer behavior tightly, and perceived close-peer behavior has not shifted.

Figure 2: Share of Australian 14–15-year-olds' own use by close friend count of believed users



Notes: Figure 2 plots the nonparametric relationship between perceived peer use of banned platforms and own past-week use among Australian 14–15-year-olds subject to the ban. The horizontal axis is the respondent's belief about how many of their closest peers still use a banned platform; the vertical axis is whether the respondent themselves reports past-week use. The relationship is descriptive: peer effects, social projection, and homophily are each consistent with the positive slope, and the cross-sectional design cannot distinguish them.

Wellbeing While compliers report less pressure to check their phones and more quality time with family and friends, our survey also shows that compliers find it harder to keep up with friends and report feeling more bored. These patterns are consistent with compliance carrying a significant social cost. Item-level results and a longer discussion appear in Appendix A3.

Ultimately, for goods that are highly social, it is unlikely that modest changes to the cost of access or personal sanctions will shift compliance unless the peer environment changes with them. Our evidence in this section suggests the social environment has not moved in Australia. We next ask what share of peers teens would require before complying themselves, and what this implies for whether higher compliance could become self-sustaining.

5 Compliance Thresholds and Norms

Whether compliance can consolidate depends on the two margins introduced above. The first is the share of peers each teen requires before complying themselves, and how observed compliance compares to that threshold. The second is whether compliance carries the social status that would sustain the high-compliance state if it were reached. We address both directly through a mechanism survey.

Sample We surveyed 239 Australian teenagers aged 13–18 in April 2026. Respondent characteristics and recruitment details are discussed in Appendix A2.

Design Respondents report what share of peers their age they believe are currently complying with the ban (the *prior*). Second, they state the share of peers that would need to comply before they themselves would (the *threshold*); the reference group is randomized across respondents among age peers, classmates in their own grade, students at their own school, and "a typical person your age," allowing us to test whether the threshold depends on how the peer group is drawn.[3]

Priors and thresholds Among banned 14–15-year-olds, teens believe on average that 30% of peers their age have complied, close to the 27% compliance rate implied by the main survey. Stated thresholds, by contrast, cluster near the top of the distribution. On average, banned teens require 69% of their peers to comply before complying themselves. This figure is robust across reference groups (62–69% across the age-peer, grade, school, and typical-other framings). Asked how many peers a typical person their age would need to see comply before that person would, banned respondents answer 69% on average—essentially the same share they require for themselves. Taken together, priors on peer compliance are low and required thresholds are high.

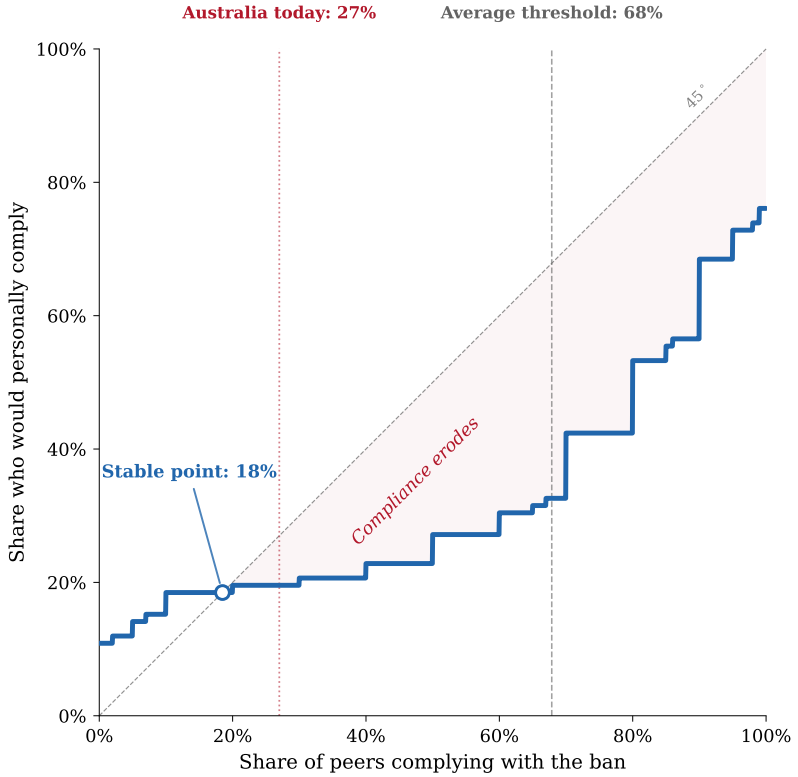
Current compliers are informative as a group on its own. Their prior belief about aggregate compliance is 44% (a slight overestimation), consistent with social forces being one reason for initial compliance alongside other mechanisms such as account deletion. Their own threshold is 63%, suggesting compliers may not remain compliant if peer behavior does not move with them. To probe this, we ask compliers whether they would continue to comply if the rate of compliance were 33%; about 50% of them would stop. The resulting compliance, for this sample, would be on the order of 18%.

The equilibrium view In Appendix A1 we develop a simple model of compliance under network effects, in the spirit of Granovetter (1978). In such a model, equilibria are points where expected and realized compliance coincide: the share of teens willing to comply, given their belief about peer behavior, equals the share who actually comply. Figure 3 plots the empirical counterpart of this idea: the CDF of stated thresholds, $F(x)$. Intuitively, $F(x)$ is the share of teens who would comply if peer compliance were at x or higher: where $F(x)$ exceeds x, more teens want to comply than currently do, so compliance rises, and where $F(x)$ falls short, compliance erodes. Equilibria lie at intersections with the 45-degree line, and their stability is determined by whether F rises more slowly (stable) or more quickly (unstable) than the line at the crossing.

The picture is stark. F lies below the 45-degree line across almost the entire unit interval, crossing it once near the origin at roughly 18%—the share implied by the unraveling exercise above. Remarkably, this parsimonious model of threshold compliance produces an equilibrium close to, but below, the 27% rate currently observed. Australia's observed compliance lies squarely in the region where $F(x) < x$, the shaded band in the figure: at this peer-compliance level, the share of respondents whose threshold has been reached is smaller than the share that would need to comply to sustain it. Compliance is therefore more likely to erode than to rise.

[3]The threshold elicitation is naturalistic: it asks respondents about a decision they plausibly already deliberate over, using reference categories (age peers, classmates in their grade, students at their school, and a typical person their age) through which adolescents organize their social environment. Moreover, the question does not send a clear signal regarding experimenter expectations (de Quidt et al., 2018; Haaland et al., 2023).

Figure 3: Empirical threshold distribution and the equilibrium set, banned 14–15-year-olds

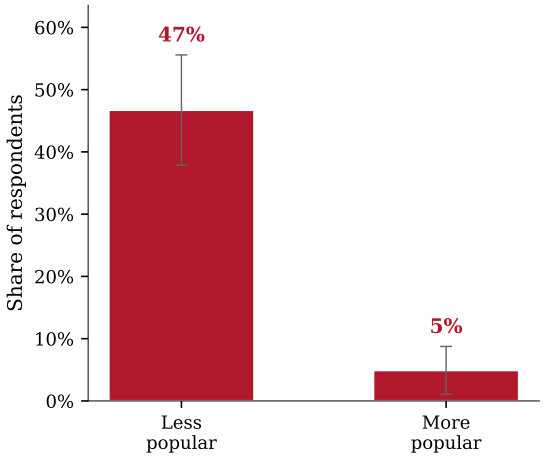


Notes: Figure 3 plots the empirical CDF of stated compliance thresholds among Australian teenagers. For each peer-compliance level x on the horizontal axis, the step function $F(x)$ gives the share of respondents whose stated threshold—the share of peers required for the respondent personally to comply—is at or below x. Equilibria of the coordination game of Appendix A1 lie at intersections with the 45-degree line, $x = F(x)$; stability is determined by the slope of F at the crossing. F lies strictly below the 45-degree line across almost the entire unit interval, crossing it once near the origin: the only stable compliance level is roughly 18%. Australia's observed compliance among 14–15-year-olds (27%) sits in the shaded region where $F(x) < x$, under which individually optimal behavior pushes compliance downward.

Sustaining an equilibrium Why does a high-compliance equilibrium seem unsustainable? The lack of a common, coordinated replacement for important functions of social media is a possible explanation. Another one has to do with *who* complies with the ban. When popular, influential teens stay on the platforms, being on social media remains the cool thing to do, and leaving carries a social cost. When the same teens leave, the meaning flips: staying behind starts to look uncool, and compliance becomes attractive in its own right. Whether compliance can grow therefore depends not only on how many teens comply, but on who they are: the same population of compliance thresholds can support a high or low equilibrium depending on the status profile of early compliers (Appendix A1; Figure A2). This is the second margin from the introduction, and it operates whether or not the tipping point is crossed.

Whether the threshold distribution could ever shift to support a high-compliance equilibrium depends on the status profile of the teens who continue to use. If non-compliance remains concentrated among more popular teens, continued use may keep exerting a social pull even after aggregate compliance rises. This would be the opposite of the cigarette precedent, in which higher-status smokers complied first and connected groups quit in concert, with smoking progressively

Figure 4: Perceived popularity of compliers with the Australian social media ban



Notes: Figure 4 reports the share of Australian teenagers who say teenagers who comply with the ban are less or more popular among their peers than those who do not. The third response category ("equally popular") is reported in the body but omitted here. Error bars represent 95% confidence intervals.

concentrated among lower-status groups and continuing smokers becoming peripheral in their social networks (Christakis and Fowler, 2008; Hiscock et al., 2012). A ban is therefore more likely to be effective if it not only raises compliance through direct enforcement, but also creates a low social status signal for non-compliers.

Our status evidence points in the opposite direction required for a high compliance. As shown in Figure 4, 47% of respondents say compliers with the ban are *less* popular, 48% say equally popular, and only 5% say more popular.[4] Among current users of banned platforms, the share saying compliers are less popular rises to 52%. Compliance has therefore not become the high-status behavior, and non-compliance has not become socially stigmatized. The teens whose continued use carries the most social weight are the ones still on the platforms, holding required thresholds high and the equilibrium low. This is exactly the margin on which complementary interventions—public-health campaigns, school-level visibility of non-use, and other tools that make non-use observable and socially validated—would need to operate. We return to these design implications in the discussion.

6 Discussion

Four months into the Australian ban, observed compliance among 14–15-year-olds—roughly 27%—sits well below the roughly two-thirds share respondents name as their own "tipping point" threshold for complying personally. We now further discuss the interpretation of our findings and consider policy implications.

[4]Suggestive evidence in the mechanism survey is consistent with this perception. Among banned 14–15-year-olds, current users of banned platforms (non-compliers) report roughly twice the Instagram follower count of non-users on average (470 vs. 200).

Interpretation An alternative interpretation is that four months is too short a window to see the full effects: the threshold distribution F is not fixed, and as the ban persists it will bend toward shapes that support a higher stable equilibrium. This could occur under two mechanisms and the data speak differently to each.

The first version is informational: teens may not yet have learned how much the peer environment has moved. Once beliefs catch up to behavior, compliance will increase. The data rule this out directly: perceptions about peer compliance among banned 14–15-year-olds in our April wave average 30%, close to the 27% rate our main survey implies and close to the 30–40% range the eSafety, Molly Rose, and YouGov surveys report independently (eSafety Commissioner, 2026; Molly Rose Foundation and YouthInsight, 2026; YouGov, 2026). Therefore, beliefs are not systematically pessimistic about peer compliance. If anything, they are mildly optimistic.

The second version is structural. Preferences, network effects, or the moral cost of violating the ban will themselves shift as the law persists—the expressive channel through which legal rules will shift norms over long horizons (Sunstein, 1996; Bénabou and Tirole, 2026). A complementary view from the literature on norm dynamics is that norms persist when the informational environment supporting them is stable and shift when that environment changes (Giuliano and Nunn, 2021; Gelfand et al., 2024); this places the burden on what teens observe their peers doing, not only on how long the ban remains in force. We cannot rule this out, and the cigarette precedent to which we return below is a reminder that decades-scale norm change is possible. What the data can say is that the current enforcement architecture is not obviously suited to activate the expressive channel. Norm change through the expressive channel requires the rule to be visible at the level of individual behavior. In the terminology of Cialdini et al. (1990) and Bicchieri (2006, 2017), the law states an injunctive norm—what ought to be done—but teenagers observe a descriptive norm—what peers actually do, and absent visible sanctions the normative expectations that would sustain compliance as a social norm have not formed. When visible peer behavior continues to signal widespread use, the descriptive norm works against the legal message rather than reinforcing it. The Australian design places enforcement entirely upstream of the user, as previously discussed. A ban enforced on the supply side could shift behavior through account removal but offers the expressive channel little surface on which to operate.

Policy preferences Teenagers and parents disagree about both the policy and the goal of social media regulation. When asked which policy they prefer, 72% of under-16s would prefer a self-limiting application—one that allows use with built-in time controls—to an outright ban, a stated preference for a nudge over a mandate in the terminology of choice architecture (Thaler and Sunstein, 2008). When considering the goal of removing access to social media altogether, evidence from a parallel US survey—a February 2026 partnership with Gallup and the Walton Family Foundation that surveyed a nationally representative sample of nearly 2,000 adolescents aged 12–18 along with a parent in the same household (Gallup and Duckworth, 2026)—finds that 72.5% of parents would prefer their child to live in a world without social media, against 31.9% of teenagers themselves. It is interesting to note that the roughly 30% of US teenagers preferring a world without social media is close to the compliance rate among Australian teenagers.

Age- versus grade-based bans The ban operates on age ranges, but adolescent peer groups are organized around classes, which may mix exact age groups. Among 14-year-olds in our April wave, 0.8 of 10 classmates in the main class are aged 16 or older; among 15-year-olds, 4.2 of 10; among 16-year-olds, 8.2; among 17-year-olds, 9.3. The jump in within-classroom exposure to openly-using unbanned peers between 14 and 15 coincides with a large jump in use inside the banned group (57.1% at 14 vs. 68.3% at 15). We cannot read this as causal, but this pattern is consistent with the peer environment influencing compliance and suggests the current ban may have failed to change the relevant peer environment by using age-based cutoffs.

Cohort dynamics A possibility is that the threshold distribution F shifts across cohorts rather than within them. In the future, teenagers who reach the banned ages having never held an account may form weaker attachments to the platforms, especially if popular teens do not join, and report lower thresholds than today's 14–15-year-olds. Whether this translates into a higher equilibrium may depend on the strength of *cross-cohort* network effects. If the social pull operates mostly within same-age peers, future cohorts can plausibly converge on a higher equilibrium; if it comes substantially from older teens who remain on platforms after age 16, compliance among future 14-year-olds will continue to be drawn upward, much as it is today. As discussed above, the rise in use between 14- and 15-year-olds, which coincides with a rise in exposure to unbanned teens, suggests cross-cohort effects are non-trivial in the current setting.

Enforcement architecture The Act places the enforcement burden on platforms rather than users, with civil penalties up to $49.5 AUD million for platforms and no legal sanction for minors. This design choice is understood by teens: only 22% believe there are personal consequences for violating the ban, while 47% believe the consequences apply to platforms alone—a perception aligned with how the Act is actually enforced. Non-compliance therefore carries neither a direct individual cost nor a visible social signal. This distinguishes the setting from classic cooperation environments in which decentralized punishment can sustain high-compliance behavior (Fehr and Gächter, 2000, 2002).

Policy considerations The ban has generated a degree of compliance well below what teens require of their peers before complying themselves. A mandate alone is unlikely to close that gap, at least under the current design of the Australian ban. Four complementary levers—each speaking to a friction the data identify—could either lift observed compliance above the threshold or lower the threshold itself.

First, zero-access is not the only point in the policy space. Tools that operate on the intensive margin—capping time on platforms rather than prohibiting accounts outright—target the heavy use most plausibly responsible for the harms the ban is meant to address while preserving the social value of moderate use. They therefore both lower the welfare cost of the policy and reduce the private incentive to circumvent it, consistent with teens' own preference for self-limiting applications over an outright ban (Cortesi and Gasser, 2026; Allcott et al., 2022). Given the social forces previously discussed, time limits would be more effective when adopted in coordination across

a peer group, so any given teenager does not pay the social cost of cutting back alone—a logic consistent with related evidence on social pressure in technology adoption (Bursztyn et al., 2025a).

Second, aligning the scope of the ban with grade rather than individual age would remove the within-classroom unbanned peers whose open use undermines coordination.

Third, informational and marketing campaigns, including nudges, to change norms could help increase the share of compliers, lower individual thresholds, change attitudes toward compliers, and thus make a high-compliance equilibrium sustainable. Individual incentives could also help increase the share of compliers—cash payments for verified non-use, vouchers for activities that fill the freed time, even tax credits for parents whose teens stay off the platforms. The cigarette precedent is instructive: incentives, for example, through increased taxes, operated together with decades of campaigns that effectively changed perceptions of smokers.

Fourth, removing social media without providing an alternative to fill the time requires compliers to absorb a social cost that the user–non-user gap on "harder to keep up with friends" illustrated. Policies that promote alternative peer-interaction channels—in-person coordination infrastructure, after school routines that occupy the freed time collectively—would reduce that cost.

References

Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow, "The Welfare Effects of Social Media," *American Economic Review*, 2020, *110* (3), 629–676.

_ **, Matthew Gentzkow, and Lena Song**, "Digital Addiction," *American Economic Review*, 2022, *112* (7), 2424–2463.

Bénabou, Roland and Jean Tirole, "Incentives and Prosocial Behavior," *American Economic Review*, 2006, *96* (5), 1652–1678.

Bénabou, Roland and Jean Tirole, "Laws and Norms," *Journal of Political Economy*, 2026, *134* (2), 731–772.

Bicchieri, Cristina, *The Grammar of Society: The Nature and Dynamics of Social Norms*, Cambridge: Cambridge University Press, 2006.

_ , *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*, Oxford: Oxford University Press, 2017.

Braghieri, Luca, Leonardo Bursztyn, and Jan Fasnacht, "Threshold Disclosure in Collective Decisions," NBER Working Paper 34827, National Bureau of Economic Research February 2026.

_ **, Ro'ee Levy, and Alexey Makarin**, "Social Media and Mental Health," *American Economic Review*, 2022, *112* (11), 3660–3693.

Bursztyn, Leonardo, Alex Imas, Rafael Jimínez-Durán, Aaron Leonard, and Christopher Roth, "Social Dynamics of AI Adoption," NBER Working Paper 34488, National Bureau of Economic Research November 2025.

_ **, Benjamin R. Handel, Rafael Jiménez-Durán, and Christopher Roth**, "When Product Markets Become Collective Traps: The Case of Social Media," *American Economic Review*, 2025, *115* (12), 4105–4136.

_ , **Ingar Haaland, Nicolas Röver, and Christopher Roth**, "The Social Desirability Atlas," *Journal of Political Economy Microeconomics*, 2025. Forthcoming. NBER Working Paper 33920; CESifo Working Paper 11911.

_ , **Matthew Gentzkow, Rafael Jiménez-Durán, Aaron Leonard, Filip Milojević, and Christopher Roth**, "Measuring Markets for Network Goods," NBER Working Paper 33901, National Bureau of Economic Research 2025.

Carpenter, Christopher and Carlos Dobkin, "The Minimum Legal Drinking Age and Public Health," *Journal of Economic Perspectives*, 2011, *25* (2), 133–156.

Christakis, Nicholas A. and James H. Fowler, "The Collective Dynamics of Smoking in a Large Social Network," *New England Journal of Medicine*, 2008, *358* (21), 2249–2258.

Cialdini, Robert B., Raymond R. Reno, and Carl A. Kallgren, "A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places," *Journal of Personality and Social Psychology*, 1990, *58* (6), 1015–1026.

Cortesi, Sandra and Urs Gasser, "Beyond Bans: A Design-Based Approach to Children's Online Safety," 2026. Berkman Klein Center for Internet & Society, Harvard University.

de Quidt, Jonathan, Johannes Haushofer, and Christopher Roth, "Measuring and Bounding Experimenter Demand," *American Economic Review*, 2018, *108* (11), 3266–3302.

DiNardo, John and Thomas Lemieux, "Alcohol, Marijuana, and American Youth: The Unintended Consequences of Government Regulation," *Journal of Health Economics*, 2001, *20* (6), 991–1010.

eSafety Commissioner, "Social Media Minimum Age: Compliance Update, March 2026," 2026. Australian Government, Office of the eSafety Commissioner. Accessed 2026-04-26.

Fehr, Ernst and Simon Gächter, "Cooperation and Punishment in Public Goods Experiments," *American Economic Review*, 2000, *90* (4), 980–994.

_ **and** _ , "Altruistic Punishment in Humans," *Nature*, 2002, *415* (6868), 137–140.

Gallup and Angela Duckworth, "How Parents and Children Think About Social Media," 2026. Nationally representative survey of 1,801 US parent–child dyads, fielded February 2026.

Gelfand, Michele J., Sergey Gavrilets, and Nathan Nunn, "Norm Dynamics: Interdisciplinary Perspectives on Social Norm Emergence, Persistence, and Change," *Annual Review of Psychology*, 2024, *75*, 341–378.

Giuliano, Paola and Nathan Nunn, "Understanding Cultural Persistence and Change," *Review of Economic Studies*, 2021, *88* (4), 1541–1581.

Granovetter, Mark, "Threshold Models of Collective Behavior," *American Journal of Sociology*, 1978, *83* (6), 1420–1443.

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart, "Designing Information Provision Experiments," *Journal of Economic Literature*, 2023, *61* (1), 3–40.

Haidt, Jonathan, *The Anxious Generation: How the Great Rewiring of Childhood Is Causing an Epidemic of Mental Illness*, New York: Penguin Press, 2024.

Hiscock, Rosemary, Linda Bauld, Amanda Amos, Jennifer A. Fidler, and Marcus Munafò, "Socioeconomic Status and Smoking: A Review," *Annals of the New York Academy of Sciences*, 2012, *1248* (1), 107–123.

Molly Rose Foundation and YouthInsight, "Australia's Social Media Ban: Is It Working?," 2026. Survey of 1,050 Australian children aged 12–15, fielded 12–31 March 2026. Accessed 2026-04-26.

Parliament of Australia, "Online Safety Amendment (Social Media Minimum Age) Act 2024," 2024. Commenced 10 December 2025. Penalties of up to A$49.5 million for non-compliant platforms. Accessed 2026-04-26.

Schelling, Thomas C., "Dynamic Models of Segregation," *Journal of Mathematical Sociology*, 1971, *1* (2), 143–186.

Sunstein, Cass R., "On the Expressive Function of Law," *University of Pennsylvania Law Review*, 1996, *144* (5), 2021–2053.

Thaler, Richard H. and Cass R. Sunstein, *Nudge: Improving Decisions About Health, Wealth, and Happiness*, New Haven, CT: Yale University Press, 2008.

Warner, Stanley L., "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias," *Journal of the American Statistical Association*, 1965, *60* (309), 63–69.

YouGov, "New YouGov Research Shows Cautious Optimism as Australians Assess Impact of Under-16 Social Media Ban," 2026. Online survey of 1,070 Australian adults, fielded 12–14 January 2026. Accessed 2026-04-26.

Online Appendix:
Not for publication

Our supplementary material is structured as follows. Appendix A1 outlines the model. Appendix A2 documents the two Australian survey waves and the US teen sample. Appendix A3 reports additional tables and figures. Appendix A4 describes the survey instruments.

A1 A model of collective compliance

In this section, we present a stylized model of compliance with the Australian under-16 social media ban. In such a setting with network effects in compliance in the spirit of Granovetter (1978), the compliance demand—the share of users who stay off social media—depends on the number of others who do so. The model formalizes two ideas. First, the share of teens who comply in equilibrium will typically depend on the curvature of the compliance demand function: a concave demand yields a high-compliance equilibrium, a convex demand yields a low one—holding constant the share of people who always comply and those who never do. Second, this curvature is itself an equilibrium object that depends, partly, on *who* complies. When high-influence individuals comply early, the demand bends concavely and equilibrium compliance is high; when they hold out, the demand bends convexly and equilibrium compliance is low.

A1.1 Setup

There is a unit mass of individuals indexed by i. Each agent chooses $a_i \in \{0, 1\}$, where $a_i = 1$ denotes compliance with the ban and $a_i = 0$ denotes non-compliance. Aggregate compliance is

$$x \equiv \int a_i \, di.$$

Agent i's payoff from complying, relative to not complying, is $v_i(x) = g_i(x) - \theta_i$, where g_i is continuous and strictly increasing, and $\theta_i \in \mathbb{R}$ is a scalar private type. Strict monotonicity of g_i captures positive network effects in compliance: the payoff from complying rises with the share of peers who comply. The private type θ_i is the (non-network) net cost of compliance, collapsing into a single scalar all heterogeneity in private platform value net of expected enforcement costs and moral or psychological costs of evading the age restriction.

Agent i complies whenever $g_i(x) \geq \theta_i$. Equivalently, when $x \geq \tau_i$, where

$$\tau_i \equiv g_i^{-1}(\theta_i) \tag{1}$$

is agent i's compliance threshold. Let F denote the CDF of τ_i. We refer to F as the compliance demand function: $F(t)$ is the share of individuals who comply when the peer-compliance rate equals t. The mass $F(0)$, individuals with $\tau_i \leq 0$, are "always compliers;" the mass $1 - F(1)$, individuals with $\tau_i > 1$, are "never compliers."

A1.2 Equilibrium

Under rational expectations,[5] equilibrium compliance solves the fixed-point condition

$$x = F(x). \tag{2}$$

The following assumption will help simplify the analysis to the case where there is a unique, stable equilibrium—closely mirroring our empirical results. We note, however, that more generally this setting admits more complex equilibria.

Assumption 1. *F is continuously differentiable on $[0, 1]$, with $F(0) > 0$, $F(1) < 1$, and $F' < 1$.*

[5] We assume rational expectations to prevent misperceptions from explaining low compliance. In practice, biased beliefs about peer activity could also explain low compliance, yet empirically we find little evidence of misperceptions about the aggregate compliance share.

Lemma 1 (Existence and uniqueness)**.** *Under Assumption 1, there is a unique interior equilibrium.*

Proof. F is a contraction on $[0,1]$ since $F' < 1$ on that interval. The Banach fixed-point theorem delivers a unique fixed point, while the boundary conditions ensure it is interior. □

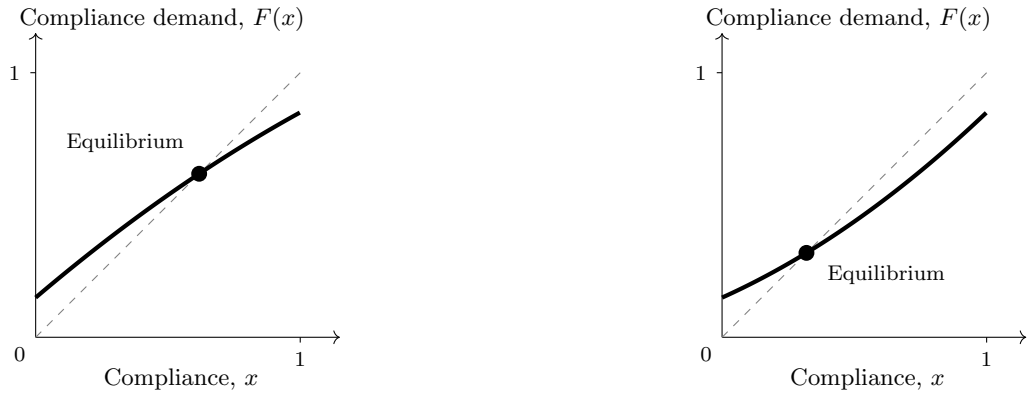
The following proposition states that, holding constant the share of those who always/never comply, the equilibrium level depends on the distribution of compliance thresholds in the interior. When the threshold distribution skews toward low values (most people need only modest peer compliance to switch), the compliance demand is concave and the equilibrium is high; when it skews toward high values (most people require near-universal peer compliance), the compliance demand is convex and the equilibrium is low.

Proposition 1 (Curvature and the location of equilibrium)**.** *Let F_C, F_V both satisfy Assumption 1, with identical boundaries $F_C(0) = F_V(0)$ and $F_C(1) = F_V(1)$. Denote by x_C^*, x_V^* their unique fixed points. If F_C is concave and F_V is convex on $[0,1]$, then $x_C^* \geq x_V^*$, with strict inequality if either curvature is strict.*

Proof. Let $L(x) \equiv F_C(0) + [F_C(1) - F_C(0)]x$ denote the line connecting the common boundary values. Concavity places F_C above this line and convexity places F_V below it, so $F_C(x) \geq F_V(x)$ (strictly on $(0,1)$ if either curvature is strict). At x_V^* this gives $F_C(x_V^*) \geq F_V(x_V^*) = x_V^*$. Since $F_C(x) - x$ is strictly decreasing and crosses zero at x_C^*, we conclude $x_C^* \geq x_V^*$ (strictly under strict curvature). □

Figure A1 illustrates the intuition. An equilibrium with high compliance requires that a substantial share of individuals have relatively low thresholds. Conversely, an equilibrium with low compliance occurs when most individual thresholds are high.

Figure A1: Curvature of compliance demand and location of equilibrium



(a) High equilibrium with concave F

(b) Low equilibrium with convex F

Notes: Figure A1 illustrates how the curvature of the compliance demand function F shapes the location of the unique equilibrium. Both panels show compliance demand functions satisfying Assumption 1 and sharing the same boundary values $F(0)$ and $F(1)$. Panel (a) shows a concave F, which yields a high-compliance equilibrium; panel (b) shows a convex F, which yields a low-compliance equilibrium. Dashed line: 45-degree line.

A1.3 Adverse selection in compliance

What drives the curvature of the compliance demand? From Equation (1) it is clear that one determinant is the distribution of idiosyncratic (non-network) net costs of compliance, θ_i. For example, if a high share of individuals are addicted to social media (Allcott et al., 2022), the

compliance demand will tend to be convex. In this section, we isolate another determinant of the curvature of compliance, which is based on the social composition of *who* complies. We argue that the same population of compliance thresholds can generate either a high- or low-compliance equilibrium depending on whether compliers tend to be high- or low-influence individuals.

Concretely, the model above treats all compliers as equally consequential for peer behavior: each contributes equally to aggregate compliance x. We now allow for heterogeneity in social influence. The goal is to microfound the curvature of the (influence-weighted) compliance demand $\tilde{F}$ that governs equilibrium, holding fixed the underlying threshold distribution F.

Setup. Each agent i is endowed with a pair (τ_i, ω_i), where τ_i is the compliance threshold and $\omega_i \geq 0$ is a social influence weight (popularity, follower count, network centrality). The pair (τ_i, ω_i) has joint density h, normalized so that $E[\omega_i] = 1$. The marginal distribution of thresholds is F. Network effects depend on the influence-weighted share of compliers,

$$\tilde{x} \equiv \int \omega_i a_i \, di.$$

Individual i complies iff $\tilde{x} \geq \tau_i$. Aggregate (unweighted) compliance is then $x = F(\tilde{x})$.

Equilibrium. The fixed-point condition becomes $\tilde{x} = \tilde{F}(\tilde{x})$,[6] where

$$\tilde{F}(t) \equiv E[\omega_i \cdot \mathbf{1}\{\tau_i \leq t\}]$$

is the influence-weighted compliance demand at peer rate t. Throughout, we assume $\tilde{F}$ satisfies Assumption 1.

In the next proposition, we conduct a thought experiment to isolate the effects of social influence. We assume that compliance thresholds are uniformly distributed and then ask how changes in the composition of social influence are reflected in the shape of the compliance demand.

Proposition 2 (Composition-driven curvature)**.** *Suppose $\tau_i \sim U[a, b]$ with $a < 0$ and $b > 1$, so F is linear on $[0, 1]$ with constant density.*[7] *Suppose further that $\omega_i = \omega(\tau_i)$, where ω is differentiable with $E[\omega_i] = 1$ and $\max_{t \in [0,1]} \omega(t) < b - a$. Then $\tilde{F}$ is convex on $[0, 1]$ if and only if ω is weakly increasing on $[0, 1]$, and concave if and only if ω is weakly decreasing on $[0, 1]$.*

Proof. Under the uniform distribution for τ_i, the compliance demand is:

$$\tilde{F}(t) = \int_a^t \omega(\tau)/(b - a) \, d\tau.$$

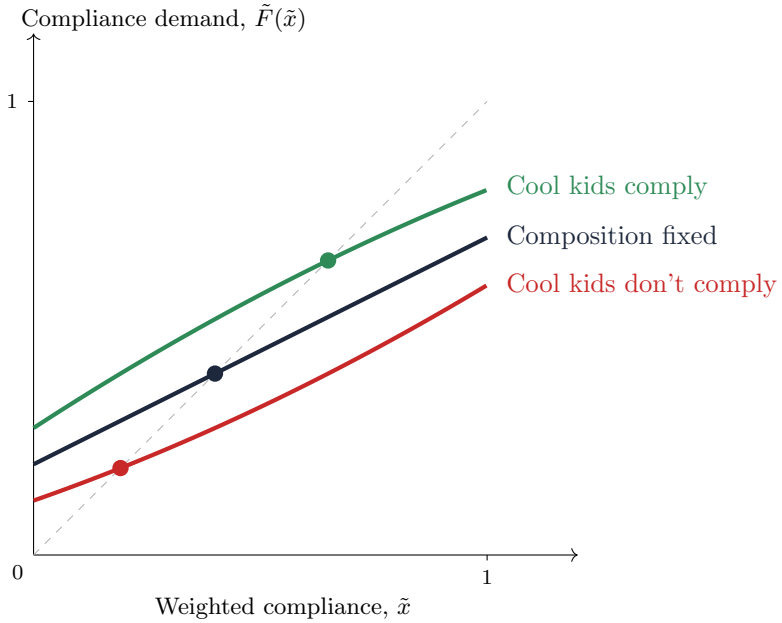
The conditions on ω ensure $\tilde{F}$ satisfies Assumption 1: $\tilde{F}(0) > 0$ and $\tilde{F}(1) < 1$ follow from $a < 0$, $b > 1$, and $\omega > 0$, while $\tilde{F}'(t) = \omega(t)/(b - a) < 1$ on $[0, 1]$. Differentiating, $\tilde{F}''(t) = \omega'(t)/(b - a)$, which has the same sign as ω'. □

[6] The policy-relevant object of interest is equilibrium (unweighted) compliance, not $\tilde{x}$. However, since F is weakly increasing, any change in composition that raises (lowers) the equilibrium influence-adjusted compliance $\tilde{x}$ also raises (lowers) actual compliance $x^* = F(\tilde{x}^*)$. The curvature results below therefore translate directly into statements about compliance.

[7] The assumption that $a < 0$ and $b > 1$ captures a setting where there is a share of always compliers and never compliers.

Figure A2 illustrates the intuition. All three curves correspond to the same population of compliance thresholds τ_i and the same baseline compliance demand F; they differ only in the joint distribution of thresholds and social influence. The black line corresponds to the case in which social influence is independent of compliance thresholds. The green line corresponds to the case in which high-influence individuals have low compliance thresholds: the early compliers are disproportionately influential, so the compliance demand is concave and the resulting compliance equilibrium is higher than baseline. This is a case of *advantageous selection* into compliance, with the distance from the black line capturing the composition effect. The red line corresponds to the opposite case, in which high-influence individuals have high compliance thresholds: the early compliers are disproportionately uninfluential and the compliance demand is convex. This is a case of *adverse selection* into compliance resulting in a lower equilibrium compliance, with the distance from the black line again capturing the composition effect.

Figure A2: Composition dependence of the compliance best-response



Notes: Figure A2 illustrates the composition dependence of the influence-weighted compliance demand $\tilde{F}$ under uniform thresholds, where the unweighted compliance demand F is linear (composition fixed). When high-influence individuals are early compliers ("cool kids comply", ω weakly decreasing in τ), $\tilde{F}$ is concave and the equilibrium social signal $\tilde{x}^*$ is high. When high-influence individuals hold out ("cool kids don't comply", ω weakly increasing in τ), $\tilde{F}$ is convex and $\tilde{x}^*$ is low. The y-intercepts differ across cases because $\tilde{F}(0) = E[\omega_i \mid \tau_i \leq 0] \cdot F(0)$ depends on composition. Dashed line: 45-degree line.

The previous proposition isolates the composition channel by assuming that thresholds are uniformly distributed. The following corollary relaxes this assumption and provides a general decomposition of the determinants of the shape of the compliance demand.

Corollary 1 (General decomposition)**.** *Suppose F has a continuously differentiable density F' on $[0,1]$ and $E[\omega_i \mid \tau_i = t]$ is continuously differentiable for t on $[0,1]$. Then for interior t,*

$$\tilde{F}''(t) = \underbrace{\frac{\partial E[\omega_i \mid \tau_i = t]}{\partial t} F'(t)}_{\text{composition channel}} + \underbrace{E[\omega_i \mid \tau_i = t] F''(t)}_{\text{preference channel}}.$$

Proof. By the law of iterated expectations, $\tilde{F}(t) = w(0)F(0) + \int_0^t m(\tau)F'(\tau)\,d\tau$, where $w(0) \equiv E[\omega_i \mid \tau_i \leq 0]$. Differentiating yields the stated decomposition. □

This corollary formalizes the two channels through which the curvature of the compliance demand is determined. The *preference channel* comes from the curvature of the unweighted compliance demand, which reflects the marginal distribution of types θ_i relative to network effects. The *composition channel* comes from the slope of average influence in the threshold of the marginal complier: whether later compliers are more or less influential than earlier ones.

A convex compliance demand with a low equilibrium, similar to the one we document empirically, can therefore arise from either channel: from right-skewed types (most individuals have high costs of compliance) or from adverse selection in compliance (high-influence individuals hold out). The two channels carry different policy implications, and it becomes essential to provide empirical evidence to understand whether both are active. Under the preference channel, intensifying enforcement uniformly shifts F up and moves the equilibrium up. Under the composition channel, uniform platform enforcement may raise average compliance but leave the social equilibrium largely unchanged if high-influence accounts remain accessible; the relevant policy lever is whether age-assurance reaches socially central users.

A2 Data and samples

Tables A1 and A2 report the composition of the three surveys used in the paper.

Table A1: Sample composition: Main Comparison Surveys

	AU 14–15 (N = 228)	AU 16–18 (N = 279)	US 14–15 (N = 168)	US 16–18 (N = 132)
Age				
Mean	14.6	17.1	14.6	16.7
% aged 14	40	0	41	0
% aged 15	60	0	59	0
% aged 16	0	30	0	44
% aged 17	0	32	0	42
% aged 18	0	38	0	14
Gender (%)				
Female	52	47	50	70
Male	46	49	47	22
Non-binary	1	2	2	8
Prefer not to say	1	2	1	1
School type (%)				
Government / public	67	55	67	76
Catholic	14	10	11	6
Independent / private	19	18	14	5
Home-schooled	0	2	8	8
Not attending	0	15	0	5
State / territory (%, AU only)				
New South Wales	48	36	–	–
Victoria	19	29	–	–
Western Australia	21	20	–	–
Queensland	6	9	–	–
South Australia	3	3	–	–
Other	3	2	–	–

Notes: Cells report within-column percentages unless otherwise labeled. Australian respondents are sampled via the survey provider Youth Insight; US respondents are sample via an online teen panel, Teen Voice. Sample sizes are indicated in brackets in each column header.

Table A2: Sample composition: Mechanism Survey

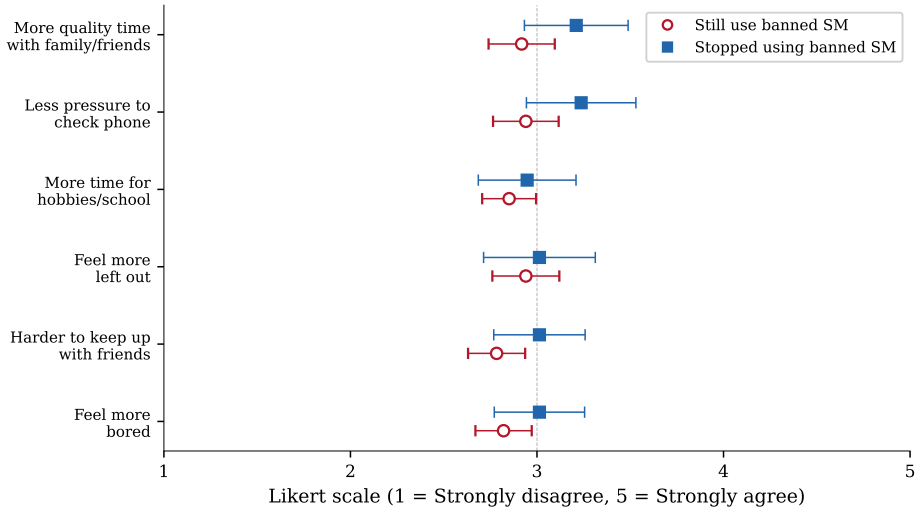
	AU (Youth Insight, 13–18) (N = 239)
Age	
Mean	16.1
% aged 13	1
% aged 14	14
% aged 15	24
% aged 16	18
% aged 17	16
% aged 18	26
Gender (%)	
Female	60
Male	38
Non-binary	1
Prefer not to say	1
School type (%)	
Government / public	56
Catholic	13
Independent / private	20
Home-schooled	2
Not attending	8
State / territory (%)	
New South Wales	36
Victoria	25
Western Australia	17
Queensland	13
South Australia	7
Other	3

Notes: This table reports the sample composition from the mechanism survey that was fielded in April 2026 through Youth Insight and is restricted to respondents aged 13–18. Sample size is indicated in the column header.

A3 Additional figures and results

Wellbeing Figure A3 reports mean responses across six wellbeing items among banned 14–15-year-olds, split by current compliance status. We find that compliers seem to spend more quality time with family/friends and feel less pressure to check their phone since the ban. In contrast, we find that they are also find it harder to keep up with friends and feel more bored. The two remaining items are similar. We view this as suggestive evidence given that compliance is an endogenous decision. The YouGov parents' survey and the Molly Rose Foundation's youth survey report qualitatively similar patterns (Molly Rose Foundation and YouthInsight, 2026; YouGov, 2026).

Figure A3: Self-reported wellbeing among banned 14–15-year-olds, by current use status



Notes: Figure A3 shows mean Likert-scale responses (1 = strongly disagree, 5 = strongly agree) on six wellbeing items among Australian 14–15-year-olds, by current use status. Users (circles) report past-week use of a banned platform; non-users (squares) do not. Across benefit-framed items, non-users report modestly higher agreement than users. Across cost-framed items, the largest user–non-user gap sits on "harder to keep up with friends" and "feeling bored". Error bars represent 95% confidence intervals; the dashed line marks the neutral midpoint.

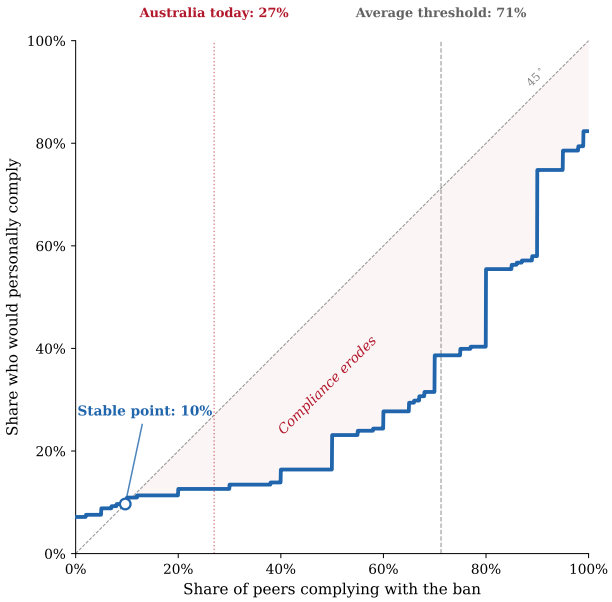
Compliance thresholds The body computes the empirical F for 13–15-year-olds. We also report the results for a more inclusive sample that pools 16–18-year-olds for whom the ban is non-binding. Table A3 reports the mean threshold, the median threshold, and the unique stable equilibrium for each sample. We also plot the empirical threshold distribution for our full sample in Figure A4. Our qualitative results remain unchanged: the only stable equilibrium is below current observed compliance. The equilibrium compliance rate moves from 10% in the pooled 13–18 sample to 18% in the body's 13–15 slice, with the mean threshold essentially unchanged.

Table A3: Threshold elicitation among banned 14–15-year-olds, by reference-group framing

Framing	N	Mean threshold (%)	Median threshold (%)	Stable equilibrium (%)
Age peers	30	62	88	30
Grade	28	62	70	18
School	32	68	80	22
Beliefs about peers	90	69	80	19

Notes: Each row corresponds to one of the four reference-group framings used in the threshold elicitation. The age-peer, grade, and school framings are randomized between subjects, so each is answered by roughly one third of the sample; the typical-other framing is asked of all respondents and is the version used to construct Figures 3 and A4. "Stable equilibrium" is the unique stable fixed point of the empirical CDF for that framing. The mean threshold sits between 62 and 69% across framings, and the stable equilibrium between 18 and 30%; the per-framing equilibria for the randomized arms ($N \approx 30$) should be read as indicative given the small cell sizes.

Figure A4: Empirical threshold distribution and the equilibrium set, ages 13–18



Notes: Figure A4 replicates Figure 3 pooling all 13–18-year-olds in the threshold wave. For each peer-compliance level x on the horizontal axis, the step function $F(x)$ gives the share of respondents whose stated threshold—the share of peers required for the respondent personally to comply—is at or below x. Equilibria of the coordination game of Appendix A1 lie at intersections with the 45-degree line, $x = F(x)$; stability is determined by the slope of F at the crossing. F lies strictly below the 45-degree line across almost the entire unit interval, crossing it once near the origin: the only stable compliance level is roughly 10%, below the body's 13–15 estimate of 18% because 16–18-year-olds—for whom the ban is non-binding—contribute thresholds at the upper end of the distribution, mechanically pulling $F(0)$ down. Australia's observed compliance among 14–15-year-olds (27%) sits in the shaded region where $F(x) < x$.

A4 Survey instruments

The paper draws on three survey waves: an Australian main survey (Youth Insight, March–April 2026; $N = 507$ aged 14–18), an Australian mechanism survey (Youth Insight, April 2026; $N = 239$ aged 13–18), and a US comparison survey (TeenVoice, March–April 2026; $N = 300$ aged 14–18). Sample composition is described in Tables A1–A2. Below we describe our survey items that produce a statistic or figure in the body. Question wording is shown verbatim in the indented blocks; response options follow each block.

A4.1 Australian main survey

The main comparison survey is the source for the compliance, circumvention, peer-belief, and wellbeing statistics in Sections 3–4 and Appendix A3.

Awareness of the ban.

> Are you aware that, since December 2025, Australian law prohibits people under 16 from holding accounts on certain social media platforms?

Options: Yes, I know about the ban and which platforms are affected; Yes, I have heard about the ban but I am not sure which platforms are affected; No, I have not heard about the ban; Not sure.

Covered platforms.

> Which of the following platforms do you believe are included in the under-16 ban? (Select all that apply.)

Fifteen options, comprising the ten covered platforms (Facebook, Instagram, Snapchat, Threads, TikTok, Twitch, X, YouTube, Kick, Reddit) and five non-covered distractors (WhatsApp, Discord, Roblox, Messenger, "I am not sure").

Perceived consequences.

> Are you aware of any consequences for under-16s who are caught using banned platforms?

Options: Yes, I am aware of consequences; No, I am not aware of any consequences; I believe the consequences are only for the platforms allowing under-16s to still use them, not for the user themselves.

Past-week use, direct.

> In the past 7 days, have you personally used any of the social media platforms included in the under-16 ban (Facebook, Instagram, Snapchat, Threads, TikTok, Twitch, X, YouTube, Kick or Reddit)?

Options: Yes / No / Prefer not to say. Conditional on Yes, respondents selected which of the ten banned platforms they had used.

Past-week use, randomized response.

> Please roll a die (physical or via a phone app). Only you can see the result — it is NOT recorded by this survey. Then follow these rules: if your number is 1, answer Yes regardless; if 6, answer No regardless; if 2, 3, 4, or 5, answer truthfully. Because we do not know what number you rolled, your individual answer is completely private.
>
> Since December 2025, Australian law prohibits people under 16 from holding accounts on certain social media platforms... In the past 7 days, have you used any social media platform that is included in the under-16 ban?

Options: Yes / No. The corrected prevalence is $\hat{p} = (\bar{y} - 1/6)/(4/6)$, where $\bar{y}$ is the share of Yes responses.

Five closest friends.

> Think about your 5 closest friends. How many of them do you think currently use a social media platform included in the under-16 ban (Facebook, Instagram, Snapchat, Threads, TikTok, Twitch, X, YouTube, Kick or Reddit)?

Options: 0 / 1 / 2 / 3 / 4 / 5 / I don't know.

Reasons for continued use (conditional on past-week use)**.**

> What are the main reasons you still use banned social media platforms? (Select all that apply.)

Options: My friends still use them and I want to stay in touch; Fear of missing out (FOMO) on what friends are doing; I use them for school-related purposes; To follow creators, celebrities, or news; Boredom or entertainment; Habit or it is hard to stop; I do not think the ban should apply to me; Other.

Awareness of circumvention methods.

> Are you aware of any ways that people your age get around the social media ban? (Select all that apply.)

Options: Using a VPN or location-changing tool; Lying about their age when signing up; Using a parent or older sibling account; Keeping an account that was created before the ban took effect; Creating a new account with a fake birthdate; Using the platform without an account (e.g., browsing without logging in); Using a different version of the platform (e.g., web browser instead of the app); I have not heard of any ways; Other.

Perceived ease of access.

> In your opinion, how easy is it for someone your age to access banned social media platforms despite the ban?

Options: Very easy / Somewhat easy / Somewhat difficult / Very difficult / Impossible / I do not know.

Personal experience with platform-side enforcement.

> Since the ban started in December 2025, have any of the following happened to you? (Select all that apply.)

Options: My account on a banned platform was removed or deactivated; I was asked to verify my age on a platform; I was blocked from creating a new account; A parent or guardian removed my access to a platform; Nothing has changed for me; I did not have accounts on any banned platforms before the ban; Other.

Wellbeing items (Figure A3)**.**

> Since the social media ban came into effect, to what extent do you agree with the following statements?
>
> - I feel less pressure to constantly check my phone.
> - I find it harder to keep up with what my friends are doing.
> - I spend more quality time with friends and family in person.
> - I feel more bored in my daily life.
> - I feel more left out of conversations and social events.
> - I have more time for hobbies, schoolwork, or other activities I enjoy.

Each item uses a 5-point Likert: Strongly disagree / Disagree / Neither agree nor disagree / Agree / Strongly agree.

Self-limiting application versus ban.

> Would you prefer to download and use an app that limits social media use instead of the ban?

Options: Yes / No.

A4.2 Australian mechanism survey

The mechanism survey is the source for the threshold elicitation, the unraveling exercise, the popularity-of-compliers item, the classroom-composition statistic, the parent-assistance item, and the Instagram-follower status proxy used in Sections 4–5 and Appendix A3. The mechanism wave also re-fielded the awareness, perceived-consequence, circumvention, reasons-for-use, and wellbeing items with identical wording to the main wave; the corresponding statistics in the body are drawn from the main wave.

Past-week use, direct. Identical wording to the main-wave direct elicitation above; used here to identify compliers and non-compliers within the threshold sample.

Prior on current peer compliance.

> Out of every 100 people your age in Australia, how many do you think have actually stopped using social media because of the ban? (Please enter a number between 0 and 100.)

Threshold elicitation. The threshold question asks how many peers would need to stop before the respondent themselves would, on a 0–100 scale. The reference group is randomized between subjects across four framings, since the framing is itself the manipulation:

> *Age peers:* Out of 100 people your age, how many would need to stop using social media for you to also stop using social media? (Please enter a number between 0 and 100.)
>
> *Grade:* Out of 100 students in your grade at your school, how many would need to stop using social media for you to also stop using social media?
>
> *School:* Out of 100 people your age at your school, how many would need to stop using social media for you to also stop using social media?
>
> *Typical other:* Now think about a typical person your age. Out of 100 people their age, how many do you think would need to stop using social media for them to also stop using social media?

The typical-other framing is asked of all respondents and is the version used to construct Figures 3 and A4; the four framings together support the cross-framing robustness check in Appendix A3.

Scenario revisions. Two follow-ups vary the assumed peer-compliance share. Current non-users are asked the low-compliance scenario:

> Now suppose that, in reality, only about one-third (33%) of people your age had stopped using social media under the policy — meaning the majority (about two-thirds, 66%) had continued to use it through workarounds. In that case, do you think you would eventually go back to using social media yourself?

Current users are asked the symmetric high-compliance scenario at 66%. Both are binary (Yes / No). The low-compliance scenario yields the unraveling rate among current compliers cited in Section 5.

Popularity of compliers.

> Do you think that teenagers who comply with the social media ban are more or less popular among their peers than those who do not comply?

Options: Less popular / Equally popular / More popular.

Classroom age composition.

> Think about the main class you attend at school (e.g., your homeroom or form class). Of every 10 classmates, roughly how many are aged 16 or older?

Options: 0–10. Used in the age- versus grade-based bans discussion in Section 6.

Parent-assisted access.

> Since December 2025, has a parent, older sibling, or other adult in your life helped you keep access to a social media platform covered by the ban?

Options: Yes / No / Prefer not to say / I have not tried to access any banned platform.

Instagram follower count.

> Approximately how many followers do you have on Instagram?

Options: 0–100 / 101–250 / 251–500 / 501–1000 / Over 1000 / Prefer not to say. Used as the status proxy in the Section 5 footnote.

A4.3 United States comparison survey

The US survey provides the same-age (14–15) and older (16–18) comparison points in Figure 1 and the alternative compliance counterfactual 35% in Section 3.

Past-week use, direct.

> In the past 7 days, have you personally used any of the following social media platforms: Facebook, Instagram, Snapchat, Threads, TikTok, Twitch, X, YouTube, Kick, or Reddit?

Options: Yes / No / Prefer not to say. Conditional on Yes, respondents selected which platforms they had used. The platform list matches the ten covered by the Australian ban so that prevalence rates are directly comparable across the two countries.