Understanding the Limits of Social Media Bans

Leonardo Bursztyn[1,2], Angela L. Duckworth[3], Rafael Jiménez-Durán[4], Aaron Leonard[1], Filip Milojević[5], Christopher Roth[6], Cass R. Sunstein[7]

[1]University of Chicago; [2]NBER; [3]University of Pennsylvania;

[4]Bocconi University, IGIER, CEPR, CESifo, and Stigler Center;

[5] Massachusetts Institute of Technology;

[6] LMU Munich, CESifo, NHH, and CEPR;

[7]Harvard Law School.

Abstract

What limits the effectiveness of social media bans? When access restrictions can be circumvented, their success may depend on the social incentives to adhere, since leaving a platform is costly when peers remain. We examine this coordination problem in Australia, which in December 2025 became the first country to ban under-16s from major social media platforms. Surveying Australian teenagers four months after implementation, we found that approximately 25% of 14–15-year-olds adhered to the ban. Yet teenagers reported requiring two out of three peers to adhere before adhering themselves, and those adhering were perceived to be less popular. These findings suggest that adherence may be difficult to sustain when it is below the level needed to make it socially attractive. Consistent with this mechanism, adherence fell to about 5% two months later. Our results highlight a limit of bans: formal rules may struggle to change behavior when social forces favor non-adherence. Policies seeking to reduce teenage social media use may therefore need to address social norms alongside platform and user incentives.

Introduction

In December 2025, Australia became the first country to ban under-16s from holding accounts on major social media platforms; similar laws are under consideration in more than a dozen countries

and in numerous U.S. states. Motivated by widespread concern that heavy social media use harms adolescent mental health[1–3], these bans target an age group in which social media use is nearly universal[4]. Yet such bans are easy to circumvent, and being off a platform is individually costly when one's peers remain on it[5]. In threshold models of collective behavior, each person adopts a behavior once the share of others who have adopted it reaches a personal threshold[6–8]. An easily circumvented ban may therefore succeed only if adherence crosses a tipping point beyond which quitting becomes self-reinforcing[9,10]. When observed adherence falls short of most teens' thresholds, threshold models predict that adherence will erode rather than grow.

Unlike tobacco or alcohol, for which enforcement operates at a physical point of sale[11,12], social media platforms can be accessed from any device at any time, and the Australian ban imposes no penalties on individual users, but imposes fines on platforms instead[13]. The ban can nonetheless influence teenagers' decisions through three channels: a personal cost (perceived sanctions on non-compliers), an effort cost (greater difficulty of access) and a social cost (a change in the peer environment that makes non-adherence to the norm expressed by the law less attractive)[14–16].

In our first survey from March and April 2026, only about a quarter of banned teens who would otherwise have used covered platforms had stopped. At the same time, teens reported needing about two-thirds of their peers to comply before complying themselves, and they perceived compliers as less popular than non-compliers. These thresholds implied that the observed level of adherence was not self-sustaining. We conducted a follow-up survey in June 2026, re-interviewing earlier respondents, and found that adherence had fallen to about one in twenty.

Our work builds on the literature on threshold models of collective behavior[7,8,17–20]. The idea that a norm collapses once enough people abandon it explains episodes ranging from the end of footbinding[21] to failed product launches[22]. We bring these models to a nationwide policy intervention aimed at changing teenagers' behavior. By combining observed adherence over time with elicited peer thresholds, we assess whether the Australian social media ban can generate a self-sustaining shift away from these platforms. Our findings show how peer interdependence can undermine a legal restriction even under active enforcement, with implications for similar

bans under consideration elsewhere. More broadly, they highlight the importance of accounting for network effects when evaluating policy interventions and government regulations on social media[23,24].

Data and Background

Australia's Online Safety Amendment (Social Media Minimum Age) Act, which came into effect in December 2025, requires designated platforms (listed in Methods) to take reasonable steps to prevent under-16s from holding accounts, with civil penalties of up to A$49.5 million[25]. The eSafety Commissioner monitored compliance from December 2025 to March 2026. In the first days after the obligation took effect on 10 December 2025, platforms removed or restricted 4.7 million accounts. By early March 2026, they had blocked more than 300,000 additional accounts[26]. On 31 March 2026, the Commissioner nonetheless announced investigations into Facebook, Instagram, Snapchat, TikTok and YouTube for potentially failing to take reasonable steps to prevent under-16s from holding accounts[27].

In our analysis, we use the results of online surveys that we conducted among Australian teenagers through an opt-in panel (Methods). The Australian main survey (March–April 2026, three to four months after the ban took effect; $N = 507$, ages 14–18) compared 14–15-year-olds, who are subject to the ban, with unbanned 16–18-year-olds. A U.S. comparison survey ($N = 300$, ages 14–18) provided a same-age benchmark for a robustness check (Supplementary Information S2). The Australian mechanism survey (April 2026; $N = 464$, ages 13–18) elicited teens' beliefs about peer adherence and their own adherence thresholds. The June 2026 follow-up survey ($N = 787$, ages 14–18) re-interviewed 422 earlier respondents and added 365 new recruits.

Results

Adherence four months after implementation

Approximately four months after implementation, 63.8% of 14–15-year-olds reported using a banned platform in the past 7 days, compared with 87.1% of unbanned 16–18-year-olds (Fig. 1). We use the older group as a benchmark for how many banned teens would have used the platforms

without the ban (the counterfactual use rate), and define adherence as the share of those teens who did not use them. Adherence is then one minus the ratio of banned teens' actual use rate to the counterfactual use rate. This gives a rate of 27%: roughly a quarter of teens who would otherwise have used banned platforms had stopped. We refer to banned teens who reported no past-week use of a covered platform as compliers and to those who reported use as non-compliers.

Within the banned group, use rose with age (57.1% at 14 versus 68.3% at 15). Because use also rose with age among unbanned teenagers, the 16–18-year-old benchmark may have overstated counterfactual use at ages 14–15, and hence overstated adherence. Although they are not subject to the ban, 16–18-year-olds may have reduced their use because of household rules, friendships with younger teens or the norm expressed by the law. These spillovers would bias our adherence estimate downward, partly offsetting the upward bias from the age gradient. As a robustness exercise, we used a same-age U.S. benchmark instead, which yields an estimated adherence rate of 35% (Supplementary Information S2).

Self-reports could understate use if teens feared admitting a banned behavior. To address this concern, we included a randomized-response module[28], which guarantees individual deniability while recovering group prevalence. It implied that 64.5% of banned teens had used a covered platform, essentially identical to the direct report. Our estimates are also in line with contemporaneous surveys by the eSafety Commissioner, the Molly Rose Foundation and Pureprofile, in which between 61% and 78% of under-16s kept an account on, or access to, banned platforms[29–31] (Supplementary Information S2).

Further, the ban appears to have redirected activity more than reduced it: reported use fell only slightly on most banned platforms, while use of messaging apps, streaming services and AI chatbots rose (Supplementary Information S2).

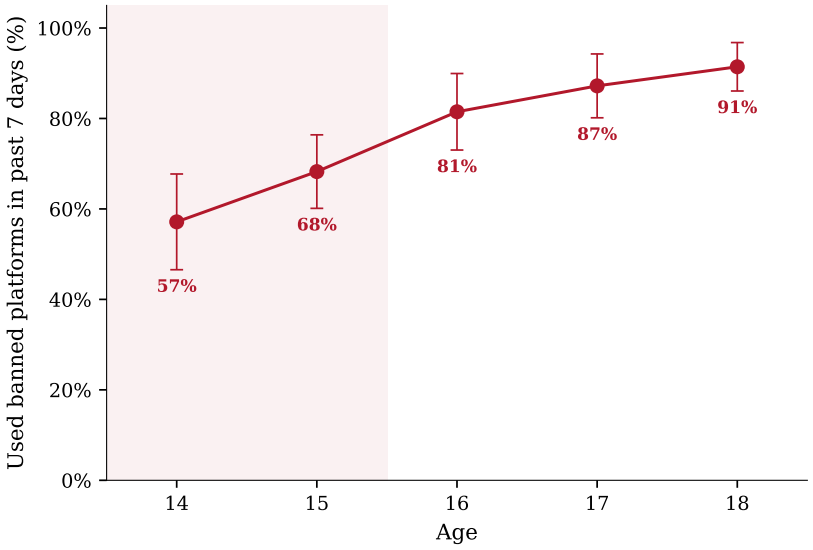


Fig. 1 | Most banned teens still used covered platforms four months into the ban. Self-reported past-7-day use of any banned platform by single year of age, Australian main survey (March–April 2026). Australian 14–15-year-olds are subject to the ban; 16–18-year-olds are not. Error bars are 95% confidence intervals. The same-age U.S. comparison appears in Supplementary Fig. 1.

Why adherence is low

To understand why adherence was low, we examine each of the three channels: personal, effort and social costs. Awareness of the ban was widespread: 98% of banned teens knew about it and 86% also knew which platforms it covered. Some also believed that non-adherence carried personal consequences: although the law imposes no sanctions on teenagers or their parents, 22% of banned teens believed that under-16s caught using banned platforms faced consequences. This belief, if anything, should have raised adherence.

While some teens believed they risked penalties, most reported little difficulty retaining access: 75% of banned teens said that getting around the restrictions was easy or very easy. This was partly because platforms had not identified many users under 16: 60% of 14–15-year-olds in our data had not had an account removed. Low-effort workarounds were also widely known among banned teens: lying about their age when signing up (57%), creating a new account with a false birthdate (44%), using a parent's or older sibling's account (42%) and using a VPN (30%). In the mechanism survey, 22% of non-compliers reported that a parent, older sibling or other adult had helped them retain access to a covered platform.

The remaining channel is social: in principle, the ban could reduce use by making non-adherence less attractive as more peers comply. At the adherence rate we observed, however, this worked in the opposite direction: with most peers still on the platforms, complying was socially costly. By respondents' own account, this cost sustained continued use: among teens still using banned platforms, 52% cited at least one social motive, most often friends' continued use or fear of missing out. Own use rose steadily with the number of close friends believed to still use banned platforms, from 15% among teens with no such friends to 86% among those with five (Supplementary Fig. 2); each additional friend was associated with a 15-percentage-point increase in the probability of using a banned platform (95% CI [11, 19]). This association reflects both influence and the tendency of friends to resemble each other, so we do not interpret it causally.

Adherence thresholds and the tipping point

To assess whether adherence could become self-sustaining, we used the mechanism survey conducted in April 2026 to measure how many peers teens would need to see quit before quitting themselves. We first asked respondents what share of peers their age they believed had stopped using social media because of the ban (the *prior*). We then asked how many out of 100 peers would need to stop before they would (the *threshold*), with the reference group randomized between respondents across age peers, students in their grade and people their age at their school. All respondents were also asked to estimate the adherence threshold of a typical teenager their age: how many out of 100 peers would need to stop using the platforms before that teenager would stop too. The elicitation concerned a decision teens plausibly already deliberated over and did not signal an expected answer[32,33].

Among banned 14–15-year-olds, teens believed on average that 30% of peers their age had complied, close to the adherence rate implied by the main survey. Stated thresholds, by contrast, clustered in the upper part of the 0–100% scale. Teens required 66% of peers to stop on average (64% among banned 14–15-year-olds). Mean thresholds were similar across the randomized reference groups (67% for age peers, 69% for grade and 63% for school; Supplementary Table 1). Asked about a typical person their age, respondents answered 73%, close to what they required

themselves.

Existing adherence looked fragile. Banned teens who had already stopped using banned platforms believed that adherence was 48%, well above the actual rate. On average they needed adherence of 47% to comply themselves, almost exactly what they believed it to be. Many compliers were therefore at the margin of complying: if they saw that more of their peers were still using the platforms, their thresholds would no longer be met. Indeed, when asked whether they would eventually resume use if only 33% of peers had stopped, about 44% said they would.

The equilibrium view. In Supplementary Information S1 we develop a model based on Granovetter's threshold model of collective behavior[7]. Equilibria are points at which the share of teens willing to comply, given peer adherence, equals the share who comply. Figure 2 plots the empirical counterpart: the cumulative distribution of stated thresholds, $F(x)$, the share of teens who would comply if peer adherence were at least x. If teens respond to the adherence they currently observe, adherence rises where $F(x) > x$ and erodes where $F(x) < x$. Equilibria lie where F crosses the 45-degree line; a crossing is stable when F cuts the line from above.

Pooling all ages, F lay below the 45-degree line across almost the entire unit interval, and the only stable adherence level was 17% (22% among banned 14–15-year-olds; Extended Data Fig. 1). Adherence observed in the main survey (27%) lay in the region where $F(x) < x$: at that level of peer adherence, fewer teens had their threshold met than were complying. The thresholds therefore implied that adherence would erode rather than rise.

Stated thresholds may differ from the thresholds that govern behavior; in Methods we report how large this discrepancy would need to be to overturn the predicted decline (Extended Data Fig. 2). The June follow-up survey provides a behavioral check on this prediction.

Who adheres

Whether the threshold distribution could shift to support high adherence depends partly on who continues to use banned social media. Unlike smoking, where connected groups quit together and continuing smokers became peripheral in their social networks[34], banned social media may retain

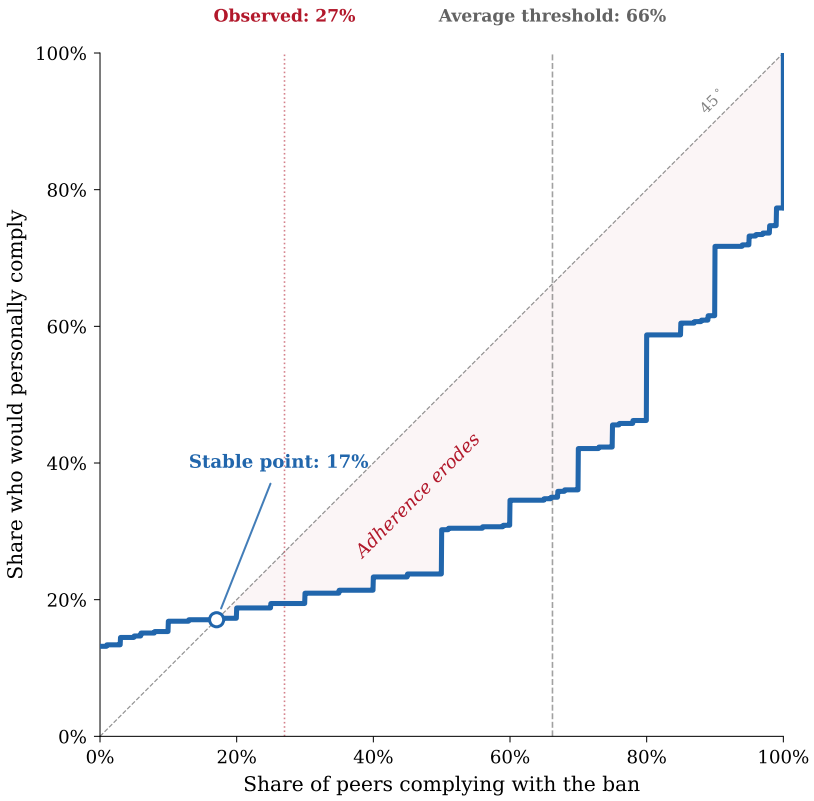


Fig. 2 | Stated adherence thresholds far exceeded observed adherence. Empirical cumulative distribution $F(x)$ of stated adherence thresholds among Australian teenagers aged 13–18, pooling the three randomized reference-group framings ($N = 463$). For each peer-adherence level x, $F(x)$ is the share of respondents whose stated threshold is at or below x. Equilibria of the model in Supplementary Information S1 lie at intersections with the 45-degree line; a crossing is stable when F cuts the line from above. The only stable adherence level is 17%. Adherence among 14–15-year-olds in the main survey (27%) sits in the shaded region where $F(x) < x$, in which adherence erodes. Extended Data Figs. 1 and 3 show banned 14–15-year-olds only and the subsample with behavior-consistent responses.

its social pull if more popular teens continue to use it even as aggregate adherence rises[35,36].

Our evidence suggests that non-adherence was concentrated among more popular teens. 45% of respondents said compliers were *less* popular than non-compliers, 51% equally popular, and only 4% more popular (Extended Data Fig. 4). Among banned 14–15-year-olds, non-compliers also reported nearly twice as many Instagram followers as compliers on average (366 vs. 201). This pattern resembles failed product launches in which awareness outpaced peer adoption, so that what spread was the decision not to switch[22]. Similarly, teens saw non-adherence, rather than adherence, as the socially rewarded behavior. In the model, such adverse selection into adherence bends the influence-weighted distribution of thresholds convexly and lowers the equilibrium (Supplementary

Information S1). Perceived status costs also appeared in the thresholds themselves: teens who saw compliers as less popular stated higher adherence thresholds (71% versus 60% on average; Extended Data Fig. 5), which lowers equilibrium adherence even without differences in influence.

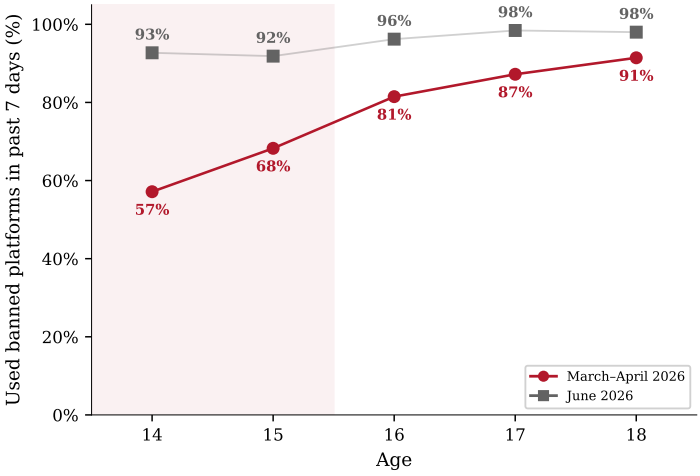
Erosion by month six

By June 2026, six months into the ban, adherence among banned 14–15-year-olds had fallen to 6%, from 27% in March–April (Fig. 3). Over the same period, the share of banned teens using a banned platform in the past week rose from 63.8% to 92.2%.

In June we surveyed 787 teenagers: 422 re-interviewed respondents from the earlier surveys (a 43% response rate) and 365 new recruits from the same panel. In total, 238 respondents were banned 14–15-year-olds, of whom 91 were new recruits. The decline in adherence was not driven by sample composition: among the 93 banned teens interviewed both in an earlier survey and in June, past-week use rose from 66% to 90%, and adherence, benchmarked against 199 unbanned teens, fell from 25% to 7% (Extended Data Fig. 6). Nor was it driven by question wording: comparing the June survey with the April mechanism survey, whose past-week use question was worded identically, gives a similar decline (Supplementary Information S2).

In June, 50% of banned teens agreed with the statement that they had stopped using one or more banned platforms when the ban began but returned "once I noticed that most people were still using them". Asked whether more or fewer of their friends were using social media than a month earlier, 63% chose "more", consistent with a self-reinforcing loop in which rising visible use makes it less likely that any teen's threshold is met. The behavior teens observed thus moved further from the norm expressed by the law[14,37]. A supply-side explanation is unlikely: enforcement tightened over this period, with investigations into five platforms under way and a government announcement in June that maximum penalties would double[27,38]. The reported return to banned platforms and the perceived increase in friends' use are consistent with the erosion of adherence implied by the April thresholds.

(a) Past-week use by age, March–April versus June 2026



(b) Use and adherence among banned teens, March–April versus June 2026

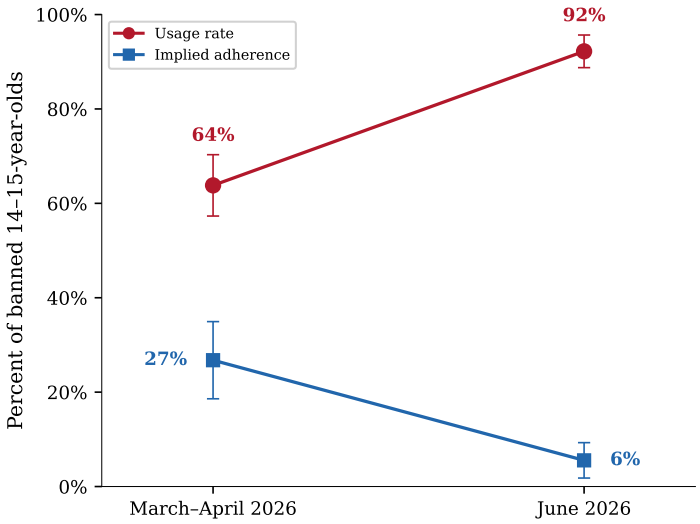


Fig. 3 | Adherence eroded between March–April and June 2026. Self-reported past-7-day use of any banned platform. April values are from the main survey (March–April 2026). (a) Use by single year of age; 14–15-year-olds are subject to the ban, 16–18-year-olds are not. (b) Red circles: past-week use among banned 14–15-year-olds. Blue squares: adherence among banned 14–15-year-olds, using unbanned 16–18-year-olds in the same survey as the counterfactual. Error bars are 95% confidence intervals.

Discussion

Six months into the Australian ban, adherence among 14–15-year-olds was roughly 6%, down from about 27% in March–April and far below the roughly two-thirds share teens said would lead them to comply. This trajectory matches what the threshold analysis implied: when adherence

starts below the tipping point, it declines.

This decline could still reverse. One possibility is that teens misperceived peer adherence and would comply more once they learned the truth; our data provide little support for this mechanism, because beliefs about peer adherence were accurate on average and adherence fell over time. Another is that preferences and norms shift as the law persists, the expressive channel through which legal rules can reshape behavior over long horizons, as with smoking[14]. Finally, the ban could prove more effective for future cohorts who reach their teens without accounts, since low adherence among teens who already used social media need not carry over to them.

If the goal is to reduce teenage social media use, four complementary levers could either lift adherence above the tipping point or lower the tipping point itself. First, coordinated limits on use, rather than prohibition, would lower both the welfare cost and the private incentive to circumvent[5,39,40]; 72% of banned teens said they would prefer an app that limits social media use over the ban. Second, norm-focused campaigns[41] and rewards for teens who can show they stayed off the platforms could coordinate a jump past the threshold. Such approaches work best when they appeal to adolescents' concerns for status and autonomy[42,43], and campaigns combined with taxes were central to reducing youth smoking[44,45].

The other two levers make adherence less costly. Third, social media occupies several hours a day, and efforts to reduce those hours are more likely to succeed if paired with attractive alternative ways to spend time. Fourth, the ban's design matters: a regime in which platforms may offer a compliant tier that removes restricted features while preserving teens' networks would build on the substitution toward messaging that teens are already making, lower teens' costs of adherence and strengthen platforms' enforcement incentives[46]. Even full adherence, however, would address only part of the problem: while evidence on the harms of social media for young people is growing, the adolescent mental health crisis also involves parents, schools and broader life circumstances[47,48].

Limitations

Sample. Our respondents are self-selected: they are teens who chose to join an online panel and take part in surveys. Compared with Australian Bureau of Statistics benchmarks[49,50], the respon-

dents in our surveys are broadly similar to the population on state of residence and the share of government-school students, and the main survey matches it on gender (49% female). They differ in that the mechanism and June surveys over-represent girls (57% and 58%) and all three under-represent Catholic-school students (11–14% versus 20%; Extended Data Table 1). Re-weighting each survey to match the population by age, gender, state and school sector left our main estimates essentially unchanged: adherence moved by at most 1 percentage point (27% to 26% in the main survey, and did not change in June) and the stable equilibrium by 1.2 points (17% to 18%).

June sample design. The June wave combined re-interviewed respondents (43% response rate) with new recruits, who were needed to maintain the size of the banned-age sample. Those who returned may differ from those who did not respond. However, the within-person estimate among 93 banned panelists yielded a magnitude similar to the cross-sectional estimate, suggesting that our results are not driven by differences in sample composition.

Stated thresholds. One limitation of our data is that our thresholds are self-reported rather than inferred from observed behavior. Two sources of error point in opposite directions. On the one hand, behavioral evidence suggests that respondents understate their thresholds: stated intentions to take costly actions exceed subsequent behavior[51], people underestimate how much others' behavior influences them[52], and from a "cold" state people underestimate the pull of cravings and social pressures they will later feel[53]. On the other hand, our threshold questions asked about stopping social media use generally, which is costlier than leaving only the banned platforms, so stated thresholds may exceed those relevant for adherence (Methods). The breakdown analysis in Methods shows that our predicted decline in adherence is robust to modest threshold overstatements, and the June follow-up provides a behavioral test: adherence fell as predicted.

Supplementary Information S3 discusses further limitations: our binary measure of use, the equal weighting of close friends and other peers, and the reading level of some questions.

References

1. Braghieri, L., Levy, R. & Makarin, A. Social media and mental health. *Am. Econ. Rev.* **112**, 3660–3693 (2022).

2. Allcott, H., Braghieri, L., Eichmeyer, S. & Gentzkow, M. The welfare effects of social media. *Am. Econ. Rev.* **110**, 629–676 (2020).

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

4. eSafety Commissioner. Digital use and risk: Online platform engagement among children aged 10 to 15. Tech. Rep., Australian Government (2025). `https://www.esafety.gov.au/research/the-online-experiences-of-children-in-australia/report-digital-use-and-risk-among-children-aged-10-to-15`.

5. Bursztyn, L., Handel, B. R., Jiménez-Durán, R. & Roth, C. When product markets become collective traps: the case of social media. *Am. Econ. Rev.* **115**, 4105–4136 (2025).

6. Schelling, T. C. Dynamic models of segregation. *J. Math. Sociol.* **1**, 143–186 (1971).

7. Granovetter, M. Threshold models of collective behavior. *Am. J. Sociol.* **83**, 1420–1443 (1978).

8. Centola, D. & Macy, M. Complex contagions and the weakness of long ties. *Am. J. Sociol.* **113**, 702–734 (2007).

9. Card, D., Mas, A. & Rothstein, J. Tipping and the dynamics of segregation. *Q. J. Econ.* **123**, 177–218 (2008).

10. Braghieri, L., Bursztyn, L. & Fasnacht, J. Threshold disclosure in collective decisions. NBER Working Paper 34827, National Bureau of Economic Research (2026). `https://ssrn.com/abstract=6246325`.

11. DiNardo, J. & Lemieux, T. Alcohol, marijuana, and American youth: the unintended consequences of government regulation. *J. Health Econ.* **20**, 991–1010 (2001).

12. Carpenter, C. & Dobkin, C. The minimum legal drinking age and public health. *J. Econ. Perspect.* **25**, 133–156 (2011).

13. eSafety Commissioner. Social media 'ban' or delay FAQs. https://www.esafety.gov.au/about-us/industry-regulation/social-media-age-restrictions/faqs (2026). Page last updated 30 September 2026.

14. Sunstein, C. R. On the expressive function of law. *Univ. Pa. Law Rev.* **144**, 2021–2053 (1996).

15. Acemoglu, D. & Jackson, M. O. Social norms and the enforcement of laws. *J. Eur. Econ. Assoc.* **15**, 245–295 (2017).

16. Bénabou, R. & Tirole, J. Laws and norms. *J. Polit. Econ.* **134**, 731–772 (2026).

17. Centola, D. The spread of behavior in an online social network experiment. *Science* **329**, 1194–1197 (2010).

18. Centola, D., Becker, J., Brackbill, D. & Baronchelli, A. Experimental evidence for tipping points in social convention. *Science* **360**, 1116–1119 (2018).

19. Guilbeault, D. & Centola, D. Topological measures for identifying and predicting the spread of complex contagions. *Nat. Commun.* **12**, 4430 (2021).

20. Guilbeault, D., Caplan, S. & Yang, C. A simple threshold captures the social learning of conventions. *Proc. Natl Acad. Sci. USA* **123**, e2508061123 (2026).

21. Mackie, G. Ending footbinding and infibulation: a convention account. *Am. Sociol. Rev.* **61**, 999–1017 (1996).

22. Centola, D. *Change: How to Make Big Things Happen* (Little, Brown Spark, New York, 2021).

23. Bursztyn, L. *et al.* Measuring markets for network goods. *Rev. Econ. Stud.* (2026). In the press.

24. Bursztyn, L., Handel, B. R., Jiménez-Durán, R. & Roth, C. Product market traps in social media: Evidence and policy implications. Policy Brief, The Hamilton Project, Brookings

Institution, Washington, DC (2026). https://static1.squarespace.com/static/5ee3119aa4c9ed2dd490b6ff/t/6a1f2369ddc69376b3ff60d3/1780425579139/20260602_THP_Handel_SocialMedia_Paper.pdf.

25. Parliament of Australia. Online Safety Amendment (Social Media Minimum Age) Act 2024 (Cth) (2024). https://www.legislation.gov.au/C2024A00127/asmade. Act No. 127, 2024.

26. eSafety Commissioner. Social media minimum age: Compliance update (March 2026). Tech. Rep., Australian Government (2026). https://www.esafety.gov.au/sites/default/files/2026-03/SocialMediaMinimumAgeComplianceUpdateMarch2026.pdf.

27. eSafety Commissioner. Five social media platforms flagged for compliance issues. https://www.esafety.gov.au/newsroom/media-releases/five-social-media-platforms-flagged-for-compliance-issues (2026). Media release, 31 March 2026.

28. Warner, S. L. Randomized response: a survey technique for eliminating evasive answer bias. *J. Am. Stat. Assoc.* **60**, 63–69 (1965).

29. eSafety Commissioner. Under the new age restrictions: Early insights from Australian parents (2026). https://www.esafety.gov.au/sites/default/files/2026-03/Under-the-new-age-restrictions-Early-insights-from-Australian-parents-March2026.pdf. Survey of 898 parents and carers. Published March 2026.

30. Molly Rose Foundation & YouthInsight. Australia's social media ban – is it working? research briefing (2026). https://mollyrosefoundation.org/wp-content/uploads/2026/04/MRF_Australia-Social-Media-Ban-Research_Briefing-April-26.pdf. Survey of 1,050 Australian children aged 12–15, fielded 12–31 March 2026. Published April 2026.

31. Pureprofile. New research: Most under-16s still use banned social media platforms six months in (2026). https://business.pureprofile.com/new-research-most-under-16s-sti

ll-use-banned-social-media-platforms-six-months-in/. Wave 2 survey of 1,025 parents, teachers and young Australians. Released 4 June 2026. Accessed 2026-10-06.

32. de Quidt, J., Haushofer, J. & Roth, C. Measuring and bounding experimenter demand. *Am. Econ. Rev.* **108**, 3266–3302 (2018).

33. Haaland, I., Roth, C. & Wohlfart, J. Designing information provision experiments. *J. Econ. Lit.* **61**, 3–40 (2023).

34. Christakis, N. A. & Fowler, J. H. The collective dynamics of smoking in a large social network. *N. Engl. J. Med.* **358**, 2249–2258 (2008).

35. Paluck, E. L. & Shepherd, H. The salience of social referents: A field experiment on collective norms and harassment behavior in a school social network. *J. Pers. Soc. Psychol.* **103**, 899–915 (2012).

36. Paluck, E. L., Shepherd, H. & Aronow, P. M. Changing climates of conflict: A social network experiment in 56 schools. *Proc. Natl Acad. Sci. USA* **113**, 566–571 (2016).

37. Tankard, M. E. & Paluck, E. L. Norm perception as a vehicle for social change. *Soc. Issues Policy Rev.* **10**, 181–211 (2016).

38. Wells, A. & Albanese, A. Stronger powers and double the penalties for world-leading social media law (2026). https://minister.infrastructure.gov.au/wells/media-release/stronger-powers-and-double-penalties-world-leading-social-media-law. Joint media release, Minister for Communications and Prime Minister of Australia, 28 June 2026.

39. Allcott, H., Gentzkow, M. & Song, L. Digital addiction. *Am. Econ. Rev.* **112**, 2424–2463 (2022).

40. Thaler, R. H. & Sunstein, C. R. *Nudge: Improving Decisions About Health, Wealth, and Happiness* (Yale University Press, New Haven, CT, 2008).

41. La Ferrara, E., Chong, A. & Duryea, S. Soap operas and fertility: evidence from Brazil. *Am. Econ. J. Appl. Econ.* **4**, 1–31 (2012).

42. Bryan, C. J. *et al.* Harnessing adolescent values to motivate healthier eating. *Proc. Natl Acad. Sci. USA* **113**, 10830–10835 (2016).

43. Yeager, D. S. *10 to 25: The Science of Motivating Young People* (Avid Reader Press / Simon & Schuster, New York, 2024).

44. Farrelly, M. C., Davis, K. C., Haviland, M. L., Messeri, P. & Healton, C. G. Evidence of a dose–response relationship between "truth" antismoking ads and youth smoking prevalence. *Am. J. Public Health* **95**, 425–431 (2005).

45. Carpenter, C. & Cook, P. J. Cigarette taxes and youth smoking: new evidence from national, state, and local Youth Risk Behavior Surveys. *J. Health Econ.* **27**, 287–299 (2008).

46. Bursztyn, L., Jiménez-Durán, R., Leonard, A., Milojević, F. & Roth, C. Aligning platform and user incentives in social media age restrictions. NBER Working Paper 35870, National Bureau of Economic Research (2026).

47. Odgers, C. L. & Jensen, M. R. Annual research review: adolescent mental health in the digital age: facts, fears, and future directions. *J. Child Psychol. Psychiatry* **61**, 336–348 (2020).

48. Odgers, C. L. The great rewiring: is social media really behind an epidemic of teenage mental illness? *Nature* **628**, 29–30 (2024).

49. Australian Bureau of Statistics. National, state and territory population, March 2026 (2026). `https://www.abs.gov.au/statistics/people/population/national-state-and-territory-population/mar-2026`. Released 17 September 2026.

50. Australian Bureau of Statistics. Schools, 2025 (2026). `https://www.abs.gov.au/statistics/people/education/schools/2025`. Released 5 March 2026.

51. Webb, T. L. & Sheeran, P. Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence. *Psychol. Bull.* **132**, 249–268 (2006).

52. Nolan, J. M., Schultz, P. W., Cialdini, R. B., Goldstein, N. J. & Griskevicius, V. Normative social influence is underdetected. *Pers. Soc. Psychol. Bull.* **34**, 913–923 (2008).

53. Loewenstein, G. Hot-cold empathy gaps and medical decision making. *Health Psychol.* **24**, S49–S56 (2005).

54. Boruch, R. F. Assuring confidentiality of responses in social research: a note on strategies. *Am. Sociol.* **6**, 308–311 (1971).

55. eSafety Commissioner. Early days, early insights: Understanding experiences of social media age restrictions at the three-month follow-up. Tech. Rep., Australian Government (2026). https://www.esafety.gov.au/sites/default/files/2026-08/Early-days-Early-Insights-SMMA-Evaluation-Understanding-experiences-of-social-media-age-restrictions-August2026.pdf.

56. YouGov. New YouGov research shows cautious optimism as Australians assess impact of under-16 social media ban (2026). https://yougov.com/articles/54334-new-yougov-research-shows-cautious-optimism-as-australians-assess-impact-of-under-16-social-media-ban. Online survey of 1,070 Australian adults, fielded 12–14 January 2026. Accessed 2026-04-26.

57. Gigerenzer, G. & Hoffrage, U. How to improve Bayesian reasoning without instruction: frequency formats. *Psychol. Rev.* **102**, 684–704 (1995).

Methods

Samples and recruitment

Australian main survey. We surveyed 507 Australian teenagers aged 14–18 between March 8 and April 6, 2026, through YouthPanel, an online panel operated by Student Edge with broad

coverage of Australian teenagers. 45% of respondents were aged 14–15 and therefore subject to the ban; 48% were male; 92% were enrolled students (Supplementary Table 2).

Australian mechanism survey. We surveyed 464 Australian teenagers aged 13–18 in April 2026 through the same panel. This survey is the source for the threshold elicitation, priors on peer adherence, the scenario questions, the popularity-of-compliers question, parent-assisted access and the Instagram-follower status proxy (Supplementary Table 3).

June follow-up survey. We surveyed $N = 787$ Australian teenagers aged 14–18 in June 2026 through the same panel: 422 re-interviewed respondents from the main and mechanism surveys (a 43% response rate among invited prior participants) and 365 new recruits. A further 5 respondents aged 13 completed the survey but were not included in the analysis. Of the 787 respondents analyzed, 238 were aged 14–15. We added new recruits because attrition alone would have left the banned-age sample too small for precise cross-sectional estimates. The June 14–15-year-old sample resembled the 14–15-year-olds in the main survey on age, gender, school sector and state (Supplementary Tables 2 and 4); within-person results for panelists interviewed in both April and June appear in Extended Data Fig. 6.

U.S. comparison survey. A U.S. comparison survey ($N = 300$, ages 14–18, fielded in March–April 2026) provides a same-age benchmark for counterfactual use; it is described in Supplementary Information S2.

Sampling frame and representativeness. All panels are opt-in, non-probability samples. Extended Data Table 1 compares each Australian survey with Australian Bureau of Statistics benchmarks. Re-weighting each survey to population margins by raking on age, gender, state and school sector changed adherence by at most 1 percentage point and the stable equilibrium by at most 1.2 points. Our key comparisons are internal, across age groups within the same panel and within person over time. Precision depends on sample size rather than population size, and we report 95%

confidence intervals throughout.

Measures

All outcomes are survey measures; verbatim wording appears in Supplementary Information S4.

Adherence. The ban covers ten platforms: Facebook, Instagram, Snapchat, Threads, TikTok, Twitch, X, YouTube, Kick and Reddit. We measured adherence using self-reported past-7-day use of any of these platforms, elicited as "Yes/No/Prefer not to say" with a platform-level follow-up. In the main survey, the question explicitly mentioned the ban. In the mechanism survey and June wave, it listed the same platforms without mentioning the ban, to limit social-desirability bias. The two versions produced comparable adherence estimates (27% in the main survey and 33% in the mechanism survey). The cross-sectional April-to-June comparison used the main survey; Supplementary Information S2 reports the comparison with the mechanism survey, whose wording matched the June survey. The within-person panel drew on both earlier surveys: 62 of the 93 banned panelists were first interviewed in the main survey, whose question named the ban, and 31 in the mechanism survey. Among the latter, for whom the wording was identical in both waves, past-week use rose from 77% to 97%. The measure captured use rather than account holding. Because the Act regulates account provision, logged-out use, such as watching YouTube without an account, falls outside its legal scope.

Adherence estimator. Let u_B be past-week use among banned 14–15-year-olds and u_C use among unbanned 16–18-year-olds in the same survey. Adherence is $1 - u_B/u_C$, the share of counterfactual users who did not use. Confidence intervals for shares are normal-approximation binomial intervals; for adherence we used the delta method, treating u_B and u_C as independent. For the panel, u_B and u_C were computed among respondents interviewed in both April and June.

Randomized response. Respondents privately rolled a die and answered Yes if it showed 1, No if it showed 6, and truthfully otherwise, the forced-response variant[54] of randomized response[28]. Because no individual answer reveals a respondent's true behavior, this method removes the incen-

tive to hide use. The corrected prevalence is $\hat{p} = (\bar{y} - 1/6)/(4/6)$, giving 64.5% against a direct report of 63.8%. Because the two estimates were nearly identical, we found no evidence that teens under-reported use when asked directly, and we used direct reports throughout.

Beliefs, thresholds and scenarios. To measure beliefs about peer adherence, we asked respondents how many out of 100 people their age had stopped using social media because of the ban. To measure adherence thresholds, we asked how many out of 100 peers would need to stop using social media for the respondent to stop too. The peer group was randomly assigned as people the respondent's age, students in their grade or people their age at their school; all respondents were also asked the same question about a typical person their age. Threshold questions referred to stopping social media use generally, whereas adherence is measured on banned platforms. Because quitting all social media is costlier than quitting the banned platforms, this difference, if anything, biases stated thresholds upward relative to the ones relevant for adherence, which works against the predicted erosion (Limitations).

To test whether stated plans respond to peer behavior, we asked current non-users whether they would eventually go back to social media if only 33% of peers had stopped, and current users whether they would stop if 66% had. Finally, the main survey asked how many of the respondent's five closest friends they believed still used a covered platform (0–5 or "I don't know").

Perceived effects of the ban. Six questions asked how teens' daily lives had changed since the ban: three on perceived social costs (keeping up with friends, feeling left out, boredom) and three on perceived benefits (less pressure to check one's phone, more in-person time, more time for other activities). Each item was scored from 1 (strongly disagree) to 5 (strongly agree) and was reported separately; we did not combine them into an index. Higher scores indicate stronger agreement with the statement, not better or worse well-being; these are single-item measures of perceived effects, not a validated well-being scale.

Other measures. The surveys also asked about awareness of the ban and of which platforms it covered, perceived consequences of using banned platforms, how easy teens thought the ban was to circumvent and which workarounds they knew of, personal experience of platform-side enforcement (including whether the respondent held an account on a banned platform before the ban) and reasons for continued use; the mechanism survey additionally asked how popular compliers were relative to non-compliers, whether a parent or other adult had helped the respondent keep access, and the respondent's approximate number of Instagram followers. Exact wording appears in Supplementary Information S4.

Instagram followers were reported in five bands (0–100, 101–250, 251–500, 501–1,000 and over 1,000). We computed mean followers by assigning each band its midpoint and a value of 1,500 to the top band, excluding respondents who preferred not to say or did not answer (68 non-compliers and 35 compliers among banned 14–15-year-olds).

Peer environment

Supplementary Fig. 2 reports use at each number of close friends believed to use banned platforms (0–5), with respondents answering "I don't know" shown separately. The continuous specification, whose slope is reported in the Results, is a linear probability model of own past-week use on the number of such friends (0–5), with age, gender and state fixed effects and robust standard errors; respondents answering "I don't know" are excluded from the regression.

Equilibrium computation

For each sample we computed the empirical cumulative distribution of stated thresholds, F, on a grid of peer-adherence levels from 0 to 100 in steps of 0.05 points, and identified stable equilibria as points at which $F(x) - x$ changes sign from non-negative to negative. Respondents who stated a threshold of 100 were treated as never complying, so $F(100)$ is below 100 and full adherence is not an equilibrium. Supplementary Information S1 gives the model and the myopic adjustment rule, $x_{t+1} = F(x_t)$, under which adherence above a stable equilibrium falls toward it.

Breakdown analysis

At observed adherence x_{obs}, the model implies erosion whenever $F(x_{\text{obs}}) < x_{\text{obs}}$. We report three measures of how wrong stated thresholds would need to be for this implication to fail. First, for any pattern of misreporting, at least $m^* = x_{\text{obs}} - F(x_{\text{obs}})$ of the population must hold true thresholds at or below x_{obs} while stating higher ones; we report m^* in percentage points and as a share of respondents who stated thresholds above x_{obs}. Second, if every true threshold were δ points below the stated one, erosion fails when $\delta \geq \delta^* = Q(x_{\text{obs}}) - x_{\text{obs}}$, where $Q(x_{\text{obs}})$ is the x_{obs}-th percentile of stated thresholds. Third, if true thresholds were a fraction λ of stated ones, erosion fails when $\lambda \leq x_{\text{obs}}/Q(x_{\text{obs}})$; we report the corresponding overstatement factor $Q(x_{\text{obs}})/x_{\text{obs}}$. Confidence intervals come from 1,000 bootstrap resamples of respondents. Extended Data Fig. 2 plots the highest stable equilibrium as a function of δ.

At the observed adherence rate of 27%, only 19% of teens reported thresholds low enough to be met. For adherence to be self-sustaining instead, at least 7.3 percentage points of the population (95% CI [0.2, 14.3]) would need to hold true thresholds at or below 27% while reporting higher ones. This amounts to 9% of teens whose stated thresholds exceeded observed adherence. Under uniform overstatement, every stated threshold would need to exceed the true one by 23 percentage points, or by a factor of 1.9.

Statistical analysis

All tests were two-sided, and we report 95% confidence intervals throughout. Confidence intervals for shares are normal-approximation binomial intervals; for adherence, a ratio of two shares, we used the delta method, treating use among banned and unbanned teens as independent. The association between own use and the number of close friends using banned platforms was estimated with a linear probability model with age, gender and state fixed effects and heteroskedasticity-robust (HC1) standard errors. The shares of respondents who rated compliers as less popular and as more popular were compared against an equal split with an exact binomial test. Confidence intervals in the breakdown analysis came from 1,000 bootstrap resamples of respondents.

Ethics

The research was approved by the University of Chicago Social and Behavioral Sciences Institutional Review Board (protocol no. IRB26-0341). Parental permission for respondents under 18 was obtained through the panel providers, and respondents gave their own informed consent at the start of each survey. AI tools were used minimally to correct grammar and spelling.

Data availability. Anonymized data are deposited at [repository and DOI] and are available to editors and referees during review.

Code availability. Code to reproduce all analyses, figures and tables is deposited at [repository and DOI].

Acknowledgements

We thank Simon Cordes, Jonathan Haidt, Ingar Haaland, Luca Henkel, Lukas Hensel, Matt Lowe, Max Müller, Shakked Noy, Aakaash Rao, Zach Rausch, and Nicolas Roever for comments. Margarita Kudaeva provided outstanding research assistance.

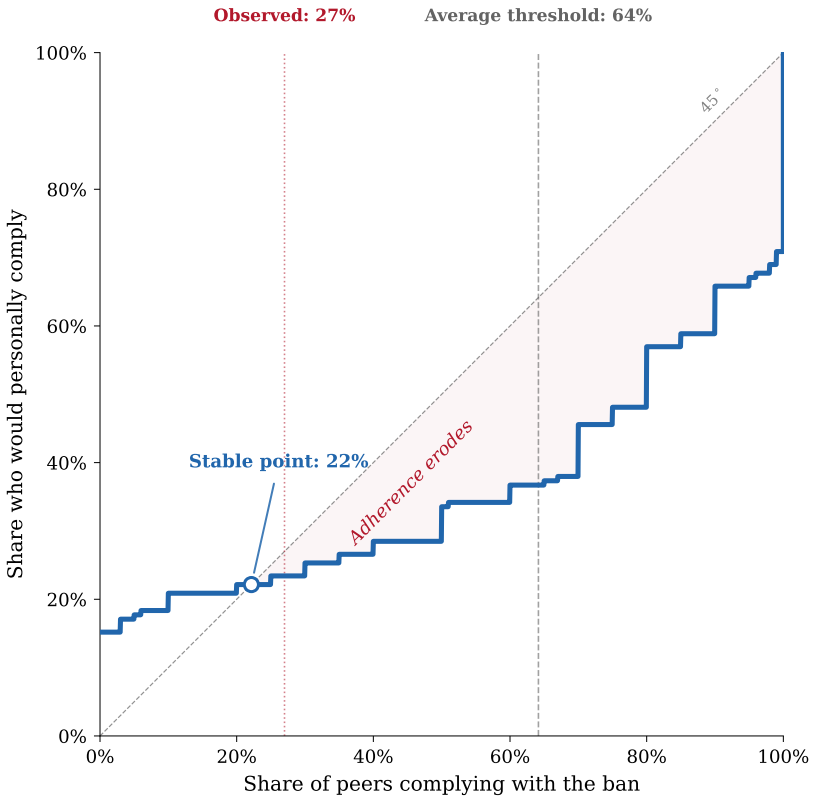
Funding. Walton Family Foundation (Angela L. Duckworth); Deutsche Forschungsgemeinschaft under Germany's Excellence Strategy, EXC 2126/2 – 390838866 (Christopher Roth).

Author contributions. All authors contributed equally to this work, including the design of the study, the analysis, and the writing of the manuscript.

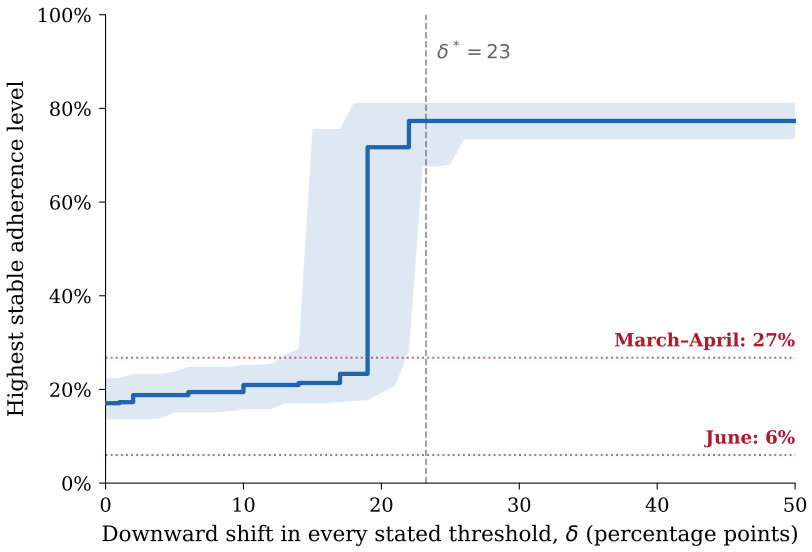
Competing interests. Leonardo Bursztyn is founder and CEO of NOMO Technologies, Inc., a company that develops digital wellness tools, and holds equity in the company. Angela L. Duckworth serves as a scientific advisor to NOMO Technologies, Inc. and holds equity in the company. NOMO had no role in the design, conduct, funding, or analysis of this research. The other authors declare no competing interests.

Additional information. Correspondence and requests for materials should be addressed to Leonardo Bursztyn (bursztyn@uchicago.edu).

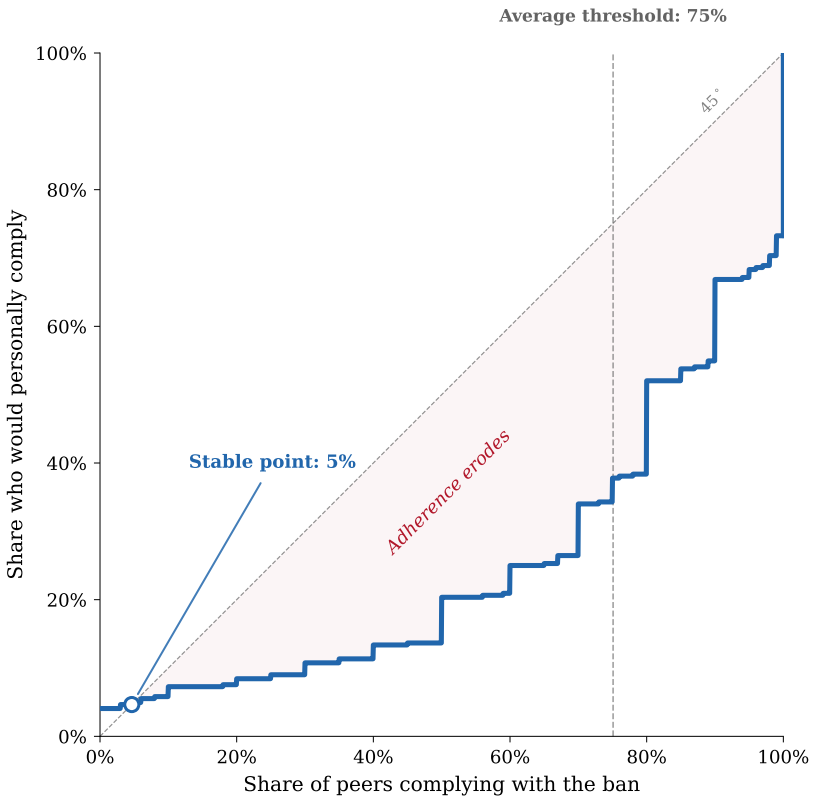
Extended Data



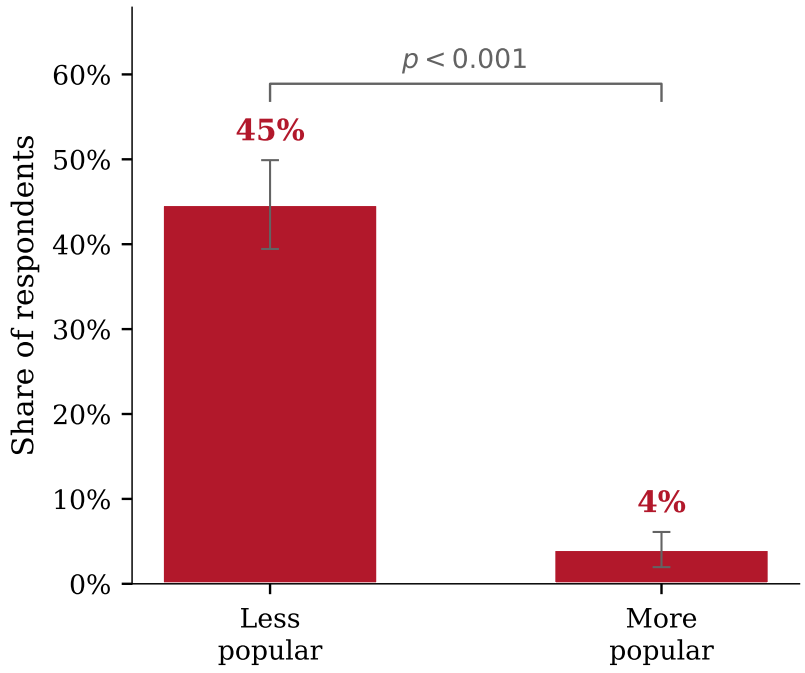
Extended Data Fig. 1 | Threshold distribution among banned 14–15-year-olds. As Fig. 2, restricted to banned 14–15-year-olds ($N = 158$). The only stable adherence level is 22%. Observed adherence among 14–15-year-olds (27%) lies in the shaded region where $F(x) < x$.



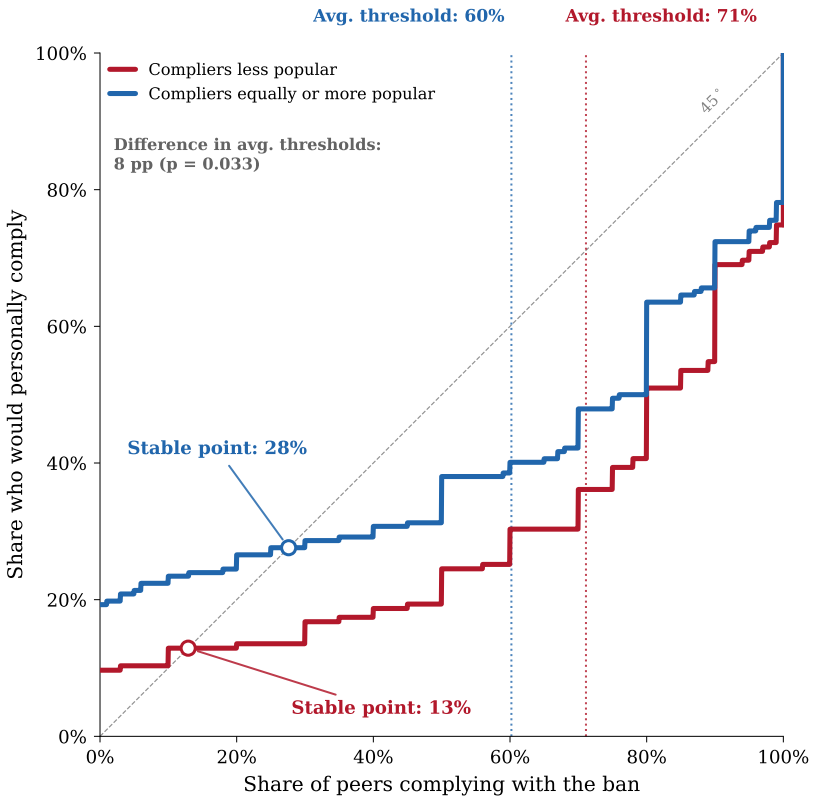
Extended Data Fig. 2 | How large must threshold misreporting be for erosion not to occur? Highest stable adherence equilibrium, pooling all respondents aged 13–18, when every stated threshold below 100 is shifted down by δ percentage points (horizontal axis). Respondents who stated a threshold of 100 are treated as never complying and are not shifted, so full adherence is not an equilibrium. Horizontal lines mark observed adherence in the main survey (March–April; 27%) and in June (6%). The dashed vertical line marks $\delta^* = 23$ points, the smallest shift at which adherence starting from the main-survey level no longer erodes. Under arbitrary heterogeneous misreporting, erosion fails only if at least 7.3 percentage points of the population hold true thresholds at or below 27% while reporting higher ones (Methods). Shaded band: 95% bootstrap confidence interval.



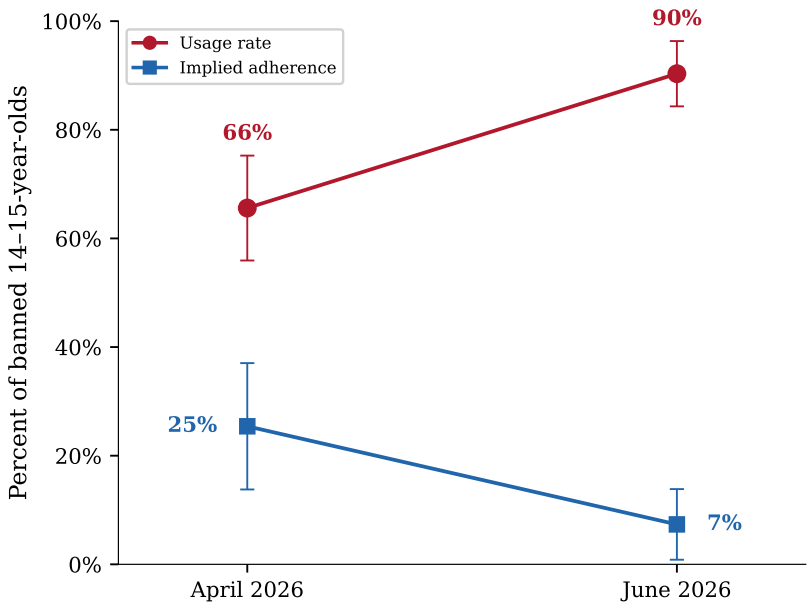
Extended Data Fig. 3 | Threshold distribution restricted to behavior-consistent responses. As Fig. 2, restricted to responses whose stated threshold is consistent with own reported use given believed peer adherence: compliers are retained only if their believed peer-adherence rate is at or above their threshold, users only if it is below. This drops roughly one quarter of responses ($N = 344$ of 463). The mean threshold rises to 75% (from 66%) and the only stable adherence level falls to 5%. Observed adherence (27%) remains in the region where $F(x) < x$.



Extended Data Fig. 4 | Teens perceived compliers as less popular than non-compliers. Share of Australian teenagers in the mechanism survey who said teenagers who comply with the ban are less or more popular among their peers than those who do not. The share answering "equally popular" (51%) is omitted. The *p*-value is from an exact two-sided binomial test of the less-popular versus more-popular split against equal shares. Error bars are 95% confidence intervals.



Extended Data Fig. 5 | Teens who saw compliers as less popular stated higher thresholds. As Fig. 2, separately for respondents who said teenagers who comply with the ban are less popular than those who do not (red; $N = 155$) and for those who said equally or more popular (blue; $N = 192$). Mean stated thresholds are 71% and 60%, respectively. The difference shown in the figure is the gap between these means after controlling for age, gender and own past-week use, with robust standard errors ($N = 336$). Stable points are computed within each group. Among banned 14–15-year-olds alone ($N = 122$), mean thresholds are 72% and 58%. The figure includes mechanism-survey responses collected from 23 April onward, when the popularity question was included. Dotted lines mark each group's average threshold.



Extended Data Fig. 6 | Within person, use rose and adherence fell between April and June. Past-week use and adherence among the 93 banned 14–15-year-olds surveyed in both April and June 2026. Red circles: share reporting past-week use of a banned platform (66% to 90%). Blue squares: adherence, using the 199 unbanned 16–18-year-olds surveyed in both April and June as the counterfactual in each month (25% to 7%). Error bars are 95% confidence intervals.

Extended Data Table 1 | Sample composition compared with Australian population benchmarks

Characteristic	Main (Mar–Apr)	Mechanism (Apr)	June	ABS benchmark
Age 14 (%)	18	14	13	20
Age 15 (%)	27	20	18	20
Age 16–18 (%)	55	65	70	60
Female (%)	49	57	58	49
NSW / VIC / QLD (%)	74	76	73	78
WA / SA / TAS / ACT / NT (%)	26	24	27	22
Government school (%)	60	58	61	63
Catholic school (%)	11	14	14	20
Independent school (%)	19	16	18	17
Headline estimates, unweighted / raked				
Adherence (%)	27 / 26	33 / 33	6 / 6	–
Stable equilibrium (%)	–	17 / 18	–	–

Columns 1–3 report the composition of the main, mechanism and June surveys; column 4 reports the corresponding population shares from the Australian Bureau of Statistics (ABS). Percentages are of all respondents, so some categories do not add up to 100: the mechanism survey also included 13-year-olds (1%), and the school rows exclude respondents who were home-schooled or not attending school (7–11%). Population shares by age, sex and state are for 31 March 2026[49]; school-sector shares are for all enrolled students in 2025[50]. Raked estimates reweight each survey to match the population by age, gender, state and school sector. Adherence is $1 - u_B/u_C$; the stable equilibrium is based on the thresholds of all respondents aged 13–18 across all three reference groups.

Supplementary Information for
Understanding the Limits of Social Media Bans

Leonardo Bursztyn*, Angela L. Duckworth, Rafael Jiménez-Durán, Aaron Leonard,
Filip Milojević, Christopher Roth, Cass R. Sunstein

*Corresponding author. Email: bursztyn@uchicago.edu

Contents: S1 A model of collective adherence; S2 Extended results; S3 Additional limitations; S4 Survey instruments; Supplementary Tables 1–6; Supplementary Figs. 1–5.

Supplementary Table 1 | Threshold elicitation among 13–18-year-olds, by reference-group framing

Framing	*N*	Mean threshold (%)	Median threshold (%)	Stable equilibrium (%)
Age peers	156	67	80	17
Grade	152	69	80	12
School	155	63	80	23
Typical person their age	463	73	80	8

Each row reports results for one version of the threshold question in the mechanism survey. Respondents were randomly assigned to one of the first three versions, which ask how many people their age, students in their grade or people their age at their school would need to stop before they would; each was answered by about one third of the sample. All respondents answered the last version, which asks the same question about a typical person their age. The stable equilibrium is the adherence level at which the distribution of thresholds crosses the 45-degree line from above (Methods). Equilibria in the first three rows rest on about 156 responses each and should be read as indicative.

Supplementary Table 2 | Sample composition: Australian main survey and U.S. comparison survey

	Australia 14–15 (N = 228)	Australia 16–18 (N = 279)	U.S. 14–15 (N = 168)	U.S. 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 (%, Australia only)				
New South Wales	48	36	–	–
Victoria	19	29	–	–
Western Australia	21	20	–	–
Queensland	6	9	–	–
South Australia	3	3	–	–
Other	3	2	–	–

Composition of the Australian main survey and the U.S. comparison survey by age group. Australian 14–15-year-olds are subject to the ban; 16–18-year-olds are not. Australian respondents were recruited through YouthPanel (Student Edge) and U.S. respondents through TeenVoice. Cells are percentages within each column, except mean age. N is the number of finished responses.

Supplementary Table 3 | Sample composition: Australian mechanism survey

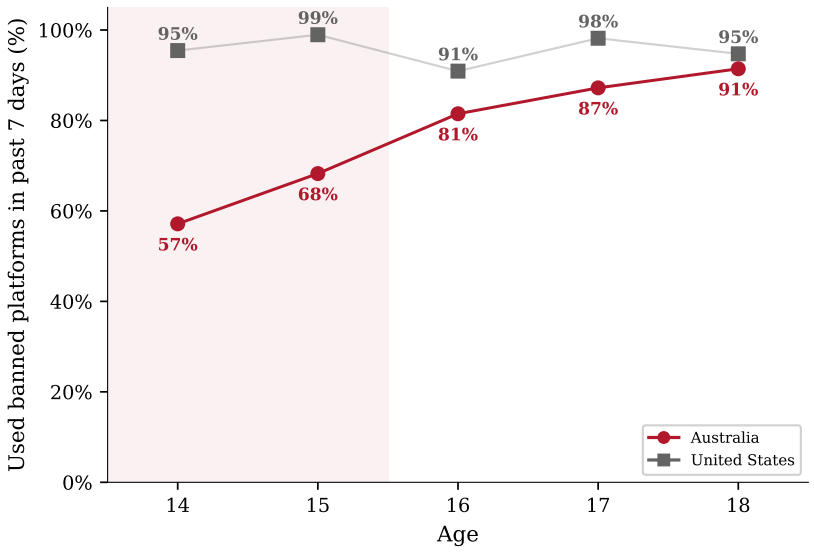
	Australia 13–18 (N = 464)
Age	
Mean	16.3
% aged 13	1
% aged 14	14
% aged 15	20
% aged 16	15
% aged 17	18
% aged 18	32
Gender (%)	
Female	57
Male	41
Non-binary	1
Prefer not to say	1
School type (%)	
Government / public	58
Catholic	14
Independent / private	16
Home-schooled	2
Not attending	10
State / territory (%)	
New South Wales	37
Victoria	27
Western Australia	14
Queensland	12
South Australia	6
Other	3

Composition of the Australian mechanism survey, fielded in April 2026 through YouthPanel (Student Edge) among teenagers aged 13–18. Cells are percentages, except mean age. N is the number of finished responses.

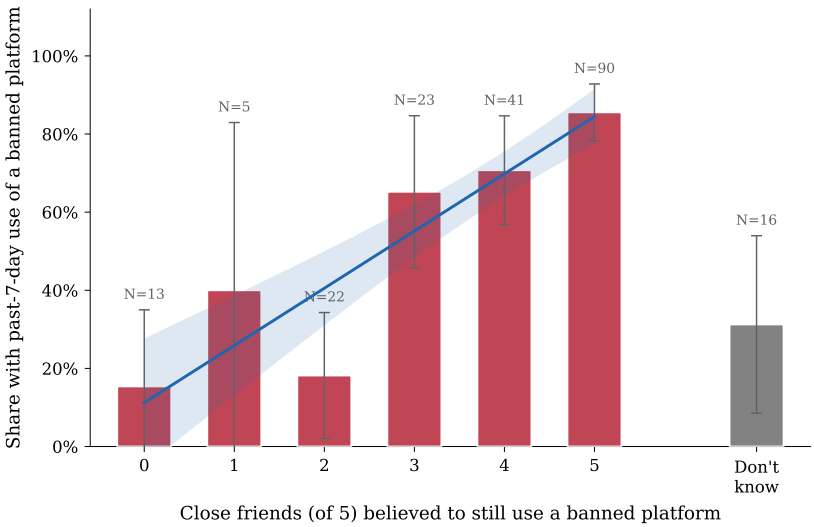
Supplementary Table 4 | Sample composition: June follow-up survey

	Australia 14–15 (N = 238)	Australia 16–18 (N = 549)
Age		
Mean	14.6	17.1
% aged 14	42	0
% aged 15	58	0
% aged 16	0	29
% aged 17	0	35
% aged 18	0	36
Gender (%)		
Female	51	61
Male	46	38
Non-binary	2	1
Prefer not to say	1	0
School type (%)		
Government / public	68	58
Catholic	12	14
Independent / private	17	19
Home-schooled	2	1
Not attending	0	8
State / territory (%)		
New South Wales	44	41
Victoria	25	24
Western Australia	19	21
Queensland	6	7
South Australia	4	5
Other	2	1

Composition of the June follow-up survey, fielded in June 2026 through YouthPanel (Student Edge). Australian 14–15-year-olds are subject to the ban; 16–18-year-olds are not. Cells are percentages within each column, except mean age. N is the number of finished responses aged 14–18; respondents aged 13 are excluded.



Supplementary Fig. 1 | Past-week use by age in Australia and the United States. Self-reported past-7-day use of any banned platform by single year of age, Australian main survey and U.S. comparison survey (March–April 2026). Australian 14–15-year-olds are subject to the ban; 16–18-year-olds are not. U.S. teenagers (gray) face no comparable policy. Use was lower at the banned ages in Australia but approximately flat across ages in the U.S.



Supplementary Fig. 2 | Use rose with the number of close friends still using banned platforms. Past-week use of a banned platform among banned 14–15-year-olds in the Australian main survey, by the number of their five closest friends believed to still use a banned platform (0–5). Bars show shares, with error bars indicating 95% confidence intervals and numbers above the bars indicating *N*. Respondents who answered "I don't know" are shown separately. The line shows a linear probability fit controlling for age, gender and state, with a 95% confidence band.

S1. A model of collective adherence

We present a stylized model of adherence to the Australian under-16 social media ban, based on Granovetter's threshold model of collective behavior[7], in which each person adopts a behavior once the share of others who have adopted it reaches a personal threshold. Here the adherence demand, the share of teens who stay off social media, depends on how many others do so. The model formalizes two ideas. First, the share of teens who comply in equilibrium depends on the curvature of the adherence demand function: holding constant the shares who always and never comply, a concave demand yields a high-adherence equilibrium and a convex demand a low one. Second, this curvature depends in part on *who* complies. When high-influence individuals comply early, demand bends concavely and equilibrium adherence is high; when they hold out, demand bends convexly and equilibrium adherence is low.

Setup

There is a unit mass of individuals indexed by i. Each agent chooses $a_i \in \{0,1\}$, where $a_i = 1$ denotes adherence to the ban and $a_i = 0$ denotes non-adherence. Aggregate adherence 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 adherence: the payoff from complying rises with the share of peers who comply. The private type θ_i is the (non-network) net cost of adherence, 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 adherence threshold. Let F denote the CDF of τ_i. We refer to F as the adherence demand function: $F(t)$ is the share of individuals who comply when the peer-adherence 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."

Equilibrium

Under rational expectations, equilibrium adherence solves the fixed-point condition

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

We assume rational expectations to prevent misperceptions from explaining low adherence; empirically, we find little evidence of misperceptions about the aggregate adherence share.

The following assumption simplifies the analysis to the case of a unique, stable equilibrium, closely mirroring our empirical results; more generally the setting admits more complex equilibria.

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

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

Proof. Because F' is continuous on the compact interval $[0,1]$ and $F' < 1$ there, $\max_{[0,1]} F' < 1$ (and $F' \geq 0$ since F is a CDF), so F is a contraction on $[0,1]$. The Banach fixed-point theorem delivers a unique fixed point, while the boundary conditions ensure it is interior. □

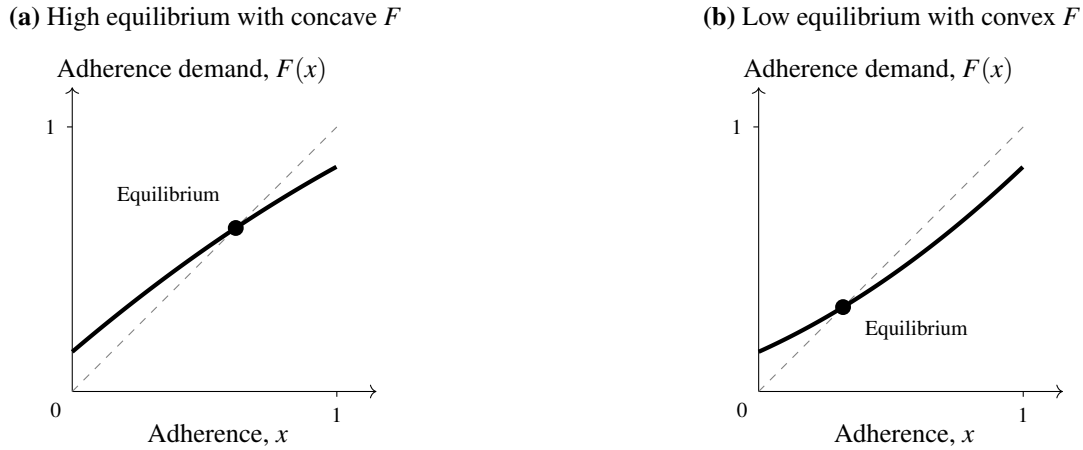
Dynamics. The fixed point is static. To interpret adherence away from equilibrium, suppose teens respond to the adherence they currently observe, so that $x_{t+1} = F(x_t)$. Under Assumption 1, F is a contraction and adherence converges to the fixed point from any starting level, falling when it starts above and rising when it starts below. For a general (empirical) F, the same rule implies that adherence between a stable crossing and the next crossing above it falls toward the stable crossing wherever $F(x) < x$. This myopic adjustment rule underlies the statement in the main text that adherence above the stable equilibrium erodes.

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

Supplementary Fig. 3 illustrates the intuition. An equilibrium with high adherence requires that a substantial share of individuals have relatively low thresholds; an equilibrium with low adherence occurs when most individual thresholds are high.



Supplementary Fig. 3 | Curvature of adherence demand and location of equilibrium. Both panels show adherence demand functions satisfying Assumption 1 and sharing the same boundary values $F(0)$ and $F(1)$. (a) A concave F yields a high-adherence equilibrium. (b) A convex F yields a low-adherence equilibrium. Dashed line: 45-degree line.

Adverse selection in adherence

What drives the curvature of the adherence demand? From Eq. (1), one determinant is the distribution of idiosyncratic (non-network) net costs of adherence, θ_i. For example, if a high share of individuals are addicted to social media[39], the adherence demand will tend to be convex. Here we isolate another determinant, based on the social composition of *who* complies: the same population of adherence thresholds can generate either a high- or low-adherence equilibrium depending on whether compliers tend to be high- or low-influence individuals.

The baseline model treats all compliers as equally consequential for peer behavior. We now allow for heterogeneity in social influence, microfounding the curvature of the influence-weighted adherence demand $\tilde{F}$ that governs equilibrium, holding fixed the underlying threshold distribution F. This extension also captures the case in which some ties, such as close friends or partners, carry more weight than others.

Setup. Each agent i is endowed with a pair (τ_i, ω_i), where τ_i is the adherence 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) adherence is then $x = F(\tilde{x})$.

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

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

is the influence-weighted adherence demand at peer rate t. Throughout, we assume $\tilde{F}$ satisfies Assumption 1. The policy-relevant object is equilibrium unweighted adherence, not $\tilde{x}$; since F is weakly increasing, any change in composition that raises (lowers) equilibrium $\tilde{x}$ also raises (lowers) actual adherence $x^* = F(\tilde{x}^*)$, so the curvature results translate directly into statements about adherence.

The next proposition conducts a thought experiment isolating the effects of social influence: adherence thresholds are uniformly distributed, and we ask how changes in the composition of social influence are reflected in the shape of the adherence 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 (the bounds capture a setting with shares of always compliers and never compliers). Suppose further that $\omega_i = \omega(\tau_i)$, where ω is differentiable and strictly positive 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 $[0,1]$.*

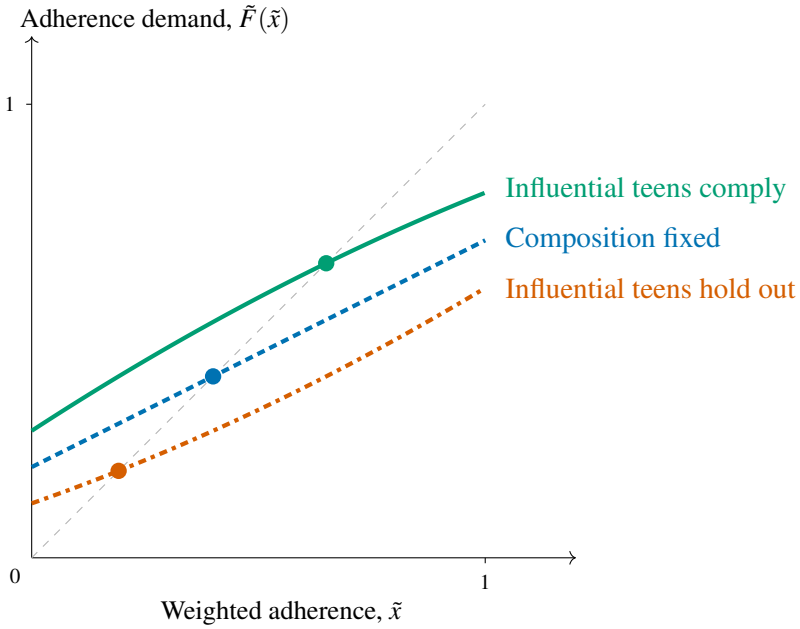
Proof. Under the uniform distribution for τ_i, the adherence demand is

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

The conditions on ω ensure $\tilde{F}$ satisfies Assumption 1: $\tilde{F}(0) > 0$ and $\tilde{F}(1) < 1$ follow from $a < 0$, $b > 1$, $\omega > 0$ and $E[\omega_i] = 1$, 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'. □

Supplementary Fig. 4 illustrates. All three curves correspond to the same population of adherence thresholds τ_i and the same baseline adherence demand F; they differ only in the joint distribution of thresholds and social influence. The dashed blue line is the case in which social influence is independent of adherence thresholds. The solid green line is the case in which

high-influence individuals have low adherence thresholds: early compliers are disproportionately influential, the adherence demand is concave, and equilibrium adherence is higher than baseline (*advantageous selection* into adherence). The dash-dotted vermilion line is the opposite case, in which high-influence individuals have high adherence thresholds: early compliers are disproportionately uninfluential, the adherence demand is convex, and equilibrium adherence is lower (*adverse selection* into adherence).



Supplementary Fig. 4 | Composition dependence of the adherence demand. Influence-weighted adherence demand $\tilde{F}$ under uniform thresholds, where the unweighted adherence demand F is linear (composition fixed; blue dashed). When high-influence individuals are early compliers (green solid; ω weakly decreasing in τ), $\tilde{F}$ is concave and the equilibrium social signal $\tilde{x}^*$ is high. When high-influence individuals hold out (vermilion dash-dotted; ω weakly increasing in τ), $\tilde{F}$ is convex and $\tilde{x}^*$ is low. Intercepts differ across cases because $\tilde{F}(0) = E[\omega_i \mid \tau_i \leq 0] \cdot F(0)$ depends on composition. Gray dashed line: 45-degree line. Colors follow the Okabe–Ito colorblind-safe palette with line styles redundant to color.

The previous proposition isolates the composition channel by assuming uniform thresholds. The following corollary relaxes this and decomposes the determinants of the shape of the adherence 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 $w(0) \equiv E[\omega_i \mid \tau_i \leq 0]$ and $m(\tau) \equiv E[\omega_i \mid \tau_i = \tau]$. Differentiating yields the stated decomposition. □

The *preference channel* comes from the curvature of the unweighted adherence demand, reflecting 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 adherence demand with a low equilibrium, similar to the one we document empirically, can arise from either channel, and the two carry different policy implications. Under the preference channel, intensifying enforcement uniformly shifts F up and moves the equilibrium up. Under the composition channel, uniform enforcement across platforms may increase average adherence while leaving the social equilibrium largely unchanged if influential users can still access their accounts. What matters for policy is whether age checks reach users who are central to their social networks.

S2. Extended results

Friend-count profile. Supplementary Fig. 2 shows past-week use among banned 14–15-year-olds at every value of the number of close friends believed to still use a banned platform. Use rises across the range; the slope of the continuous specification is reported in the main text.

Comparison with other surveys. Other surveys fielded after the ban report magnitudes similar to ours. In the eSafety Commissioner's parent survey, 64%–69% of children who had held accounts on Facebook, Instagram, Snapchat or TikTok before the ban still had them[29]; the Molly Rose Foundation reported that 61% of 12–15-year-olds who previously held accounts on restricted platforms still had access to an active account[30]; and in a Pureprofile survey of parents, teachers and young Australians six months into the ban, 78% of under-16s were still accessing banned platforms[31]. In eSafety's nationally representative evaluation, 85.9% of 10–15-year-olds used an age-restricted platform shortly before December 2025, with or without an account, and 52.4% held an account; three months later, these shares were 81.5% and 42.1%[55]. These figures cover a wider age range than our banned sample, so they do not translate directly into an adherence rate for 14–15-year-olds. Like our measure, eSafety's use figures count logged-out use, which the Act does not regulate; its account-holding figures do not.

April baseline in Fig. 3. Fig. 3 compares June with the main survey from March–April, our main estimate of adherence. The two surveys asked about past-week use slightly differently. The main-survey question named the ban ("In the past 7 days, have you personally used any of the social media platforms included in the under-16 ban…?"), whereas the June question listed the same ten platforms without mentioning it ("In the past 7 days, have you personally used any of the following social media platforms…?"; Supplementary Information S4). Using instead the April mechanism survey, whose question was worded the same as in June, gives a similar result: adherence fell from 33% to 6% and use rose from 63.9% to 92.2%. Among banned 14–15-year-olds, reported use was nearly identical under the two April wordings (63.8% in the main survey and 63.9% in the mechanism survey), so the higher April adherence in the mechanism survey reflects higher use among unbanned 16–18-year-olds, the benchmark, rather than the wording of the question.

U.S. comparison. We surveyed 300 U.S. teenagers aged 14–18 through TeenVoice during the same period as the Australian main survey. The U.S. questionnaire listed the same ten platforms. The U.S. sample provides a same-age benchmark for counterfactual use. Use there is approximately flat across ages 14–18 (Supplementary Fig. 1), probably because it is already close to 100% at the youngest ages, which leaves little room for an age gradient. Using U.S. 14–15-year-

olds (97.5% using) as the counterfactual yields adherence of 35%. Cross-country differences in platform adoption make this benchmark imperfect.

Intensive margin and substitution. In addition to measuring whether teens complied, the mechanism survey asked how their use of each banned platform and of non-banned alternatives had changed since the ban. Responses ranged from decreased a lot (-2) to increased a lot ($+2$) on a five-point scale. Most teens reported no change on most platforms (Supplementary Table 5). Where use changed, the average reported change was negative for every banned platform except TikTok, though declines were small (-0.1 to -0.4). Compliers reported declines on all banned platforms, while non-compliers reported increased TikTok use (38% reported an increase versus 10% a decrease). Outside the banned set, use rose for Spotify (53%), streaming (42%), AI chatbots (42%) and messaging apps such as WhatsApp and iMessage (39–40%). These are retrospective self-reports of change; they suggest that the ban shifted much of non-compliers' activity to other services.

The shift toward private messaging is encouraging: it lets teens stay in touch with friends without public feeds or follower counts, the features at the center of policy concern. At the same time, the growth of services outside the ban's scope shows that the list of covered services needs regular review.

The main survey also asked past-week users how often and how long they used banned platforms ($N = 81$, 184 and 158 for banned 14–15-year-olds, Australian 16–18-year-olds and U.S. 14–15-year-olds). Banned 14–15-year-old users reported lower intensity than unbanned 16–18-year-old users: 57% versus 73% near-daily use, and less time per day (Supplementary Table 6). The Australian time question refers to a day on which the respondent used the platforms, whereas the U.S. question refers to a typical day.

Supplementary Table 5 | Reported change in platform use since the ban, banned 14–15-year-olds

	N	Mean	% decreased	% same	% increased
Banned platforms					
Facebook	89	−0.4	38	48	13
Snapchat	98	−0.4	34	55	11
Kick	43	−0.4	35	53	12
Threads	51	−0.4	37	51	12
Instagram	104	−0.3	37	47	16
X	54	−0.3	30	57	13
Reddit	57	−0.3	30	51	19
Twitch	54	−0.2	30	48	22
YouTube	140	−0.1	27	55	18
TikTok	95	0.0	29	41	29
Non-banned alternatives					
Messenger	84	0.0	26	48	26
Roblox	110	0.0	25	49	25
Fortnite and other games with chat	92	+0.1	16	59	25
Discord	90	+0.2	20	48	32
Pinterest	101	+0.3	11	54	35
iMessage / SMS	140	+0.4	11	51	39
WhatsApp	121	+0.4	11	49	40
AI chatbots (ChatGPT, Character.AI, etc.)	122	+0.5	8	50	42
Streaming (Netflix, Disney+, etc.)	133	+0.5	5	53	42
Spotify	136	+0.9	5	42	53

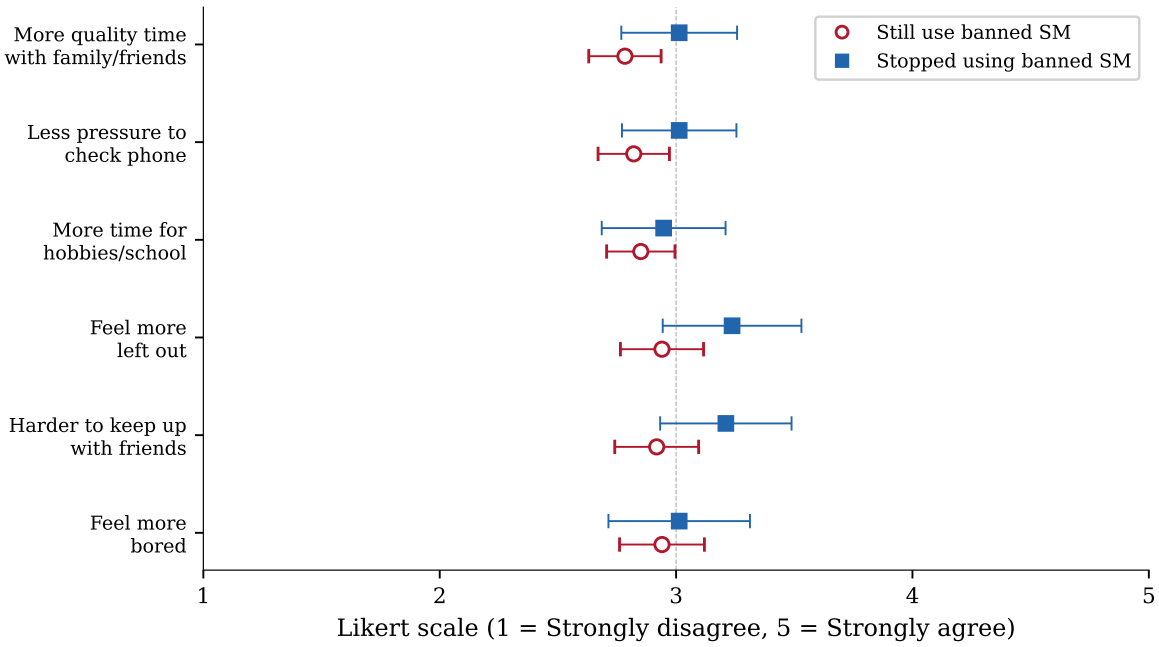
Banned 14–15-year-olds in the mechanism survey rated how their use of each platform had changed since the ban took effect, from decreased a lot (−2) to increased a lot (+2). Respondents who did not use a platform are excluded, so *N* varies by platform. Within each group, platforms are sorted by mean change. Responses are retrospective self-reports.

Supplementary Table 6 | Use intensity among past-week users

	Australia 14–15 (N = 81)	Australia 16–18 (N = 184)	U.S. 14–15 (N = 158)
Days used, past 7 days (%)			
1 day	10	4	–
2–3 days	19	10	–
4–5 days	15	12	–
6–7 days (every or almost every day)	57	73	–
Time spent per day (%)			
Less than 30 minutes	12	4	6
30 minutes to 1 hour	30	26	16
1–2 hours	32	36	25
2–3 hours	20	20	18
More than 3 hours	6	13	36

Percentages within each column, among respondents who used a banned platform in the past week. The Australian questions were asked only in the main survey and refer to a day on which the respondent used the platforms. The U.S. question refers to a typical day, and the U.S. survey did not ask on how many days respondents used the platforms. The U.S. categories above 3 hours are combined to match the Australian top category.

Perceived effects of the ban. Supplementary Fig. 5 reports responses to the six perceived-effects items among banned 14–15-year-olds, by adherence status (scoring in Methods). Point estimates suggest that compliers spent more in-person time with family and friends and felt less pressure to check their phones since the ban, but also found it harder to keep up with friends and felt more left out; none of these differences is statistically significant. The two remaining items, time for other activities and boredom, were similar across groups. The YouGov parents' survey and the Molly Rose Foundation's youth survey are broadly consistent with these patterns[30,56], as is eSafety's evaluation[55].



Supplementary Fig. 5 | Perceived effects of the ban among compliers and non-compliers. Mean responses (1 = strongly disagree, 5 = strongly agree) to six items on perceived effects of the ban among banned 14–15-year-olds in the Australian main survey. Items are reported separately, not combined into an index. Non-compliers (red circles; "Still use banned SM") reported past-week use of a banned platform; compliers (blue squares; "Stopped using banned SM") did not. Error bars are 95% confidence intervals; the dashed line marks the neutral midpoint.

Threshold robustness. Supplementary Table 1 reports mean and median thresholds and the stable equilibrium by reference-group framing. Extended Data Figs. 1 and 3 show results for banned 14–15-year-olds only and for the subsample whose responses were consistent with their behavior. In every case, the only stable equilibrium is below observed adherence.

S3. Additional limitations

Binary use measures. Our adherence measure records whether respondents used a platform in the past week, so it does not capture reductions in time spent. Questions about frequency and time spent, asked only in the main survey, suggest some reduction in use among non-compliers (Supplementary Information S2). Whether the ban reduces engagement among teens who keep using social media is an important question for future work.

Tie strength. Teens may care more about what their close friends do than about peers in general, but our elicitation asked about shares of a broad reference group and so weighted all peers equally. This simplification likely understates the pressure to use: most respondents said that most of their five closest friends still used banned platforms, so giving close friends more weight would raise the effective level of peer use teens face and reinforce the predicted decline in adherence. Supplementary Information S1 formalizes this with an influence-weighted model.

Reading level. Some questions, notably the scenario questions, used proportions and percentages that younger respondents may have found difficult to understand. The survey questions had an average Flesch–Kincaid grade level of 8.6 (ranging from 4 to 18 across questions). Threshold questions used an "out of 100 people" format, which tends to make proportions easier to understand[57].

S4. Survey instruments

The paper uses data from four surveys: the Australian main survey (YouthPanel, March–April 2026; $N = 507$, ages 14–18), the Australian mechanism survey (YouthPanel, April 2026; $N = 464$, ages 13–18), the U.S. comparison survey (TeenVoice, March–April 2026; $N = 300$, ages 14–18) and the Australian June follow-up survey (YouthPanel, June 2026; $N = 787$, ages 14–18, excluding 5 respondents aged 13). The indented blocks reproduce the exact wording of each question, with response options listed below.

Australian main survey

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), four non-covered distractors (WhatsApp, Discord, Roblox, Messenger) and "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 or restricting access on certain social media platforms including Facebook, Instagram, Snapchat, Threads, TikTok, Twitch, X, YouTube, Kick, and Reddit. In the past 7 days, have you used any social media platform that is included in the under-16 ban?

Options: Yes / No.

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.

Perceived effects of the ban.

> 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 scale: 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.

Australian mechanism survey

The mechanism survey re-fielded the awareness, perceived-consequence, circumvention, reasons-for-use and perceived-effects items with wording identical to the main survey. The past-week use item listed the same ten platforms but did not refer to the ban, to limit social-desirability bias; adherence estimates were similar to the main survey's. Statistics on awareness, costs and reasons reported in the paper are drawn from the main survey.

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 ("Which of the following platforms have you used in the past 7 days? (Select all that apply)").

We used this question to identify compliers and non-compliers in the mechanism survey and as the April comparison for the June follow-up survey.

Prior on current peer adherence.

> 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 reference group was randomized between subjects across the first three framings; the typical-other framing was asked of all respondents.

> *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?

Scenario revisions. Current non-users were asked the low-adherence 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 were asked the symmetric high-adherence scenario at 66%. Both are binary (Yes / No).

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.

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.

Australian June follow-up survey

The past-week use and peer-belief items used the same wording as the April mechanism survey, so the use item did not refer to the ban. Two retrospective items were added.

Stopped and resumed.

> Please answer whether you agree or disagree with the following statement:
>
> When the ban first took effect, I originally stopped using at least one of the banned platforms, but once I noticed that most people were still using them I went back on them.

Options: Agree / Disagree.

Change in friends' use.

> Compared to one month ago, do you think more or less of your friends are now using social media platforms (Facebook, Instagram, Snapchat, Threads, TikTok, Twitch, X, YouTube, Kick, Reddit)?

Options: More friends are using social media now compared to one month ago / Less friends are using social media now compared to one month ago.

United States comparison survey

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.