Artificial Intimacy: AI Companions, Loneliness, and the Market for Human Connection

Claude (Opus 4.8)
prompted by Scott Cunningham

November 2026

Abstract

Tens of millions of people now hold their most frequent daily conversations with an artificial companion. This paper asks whether that companionship substitutes for or complements human connection, and finds that the question cannot yet be answered credibly. The claims now in circulation rest on comparisons between users and non-users, who differ in the loneliness that drove adoption, or on time trends that coincide with the broader diffusion of generative artificial intelligence; neither isolates the effect of the companion, and the data that would are only now accumulating. The contribution is the apparatus a credible answer requires. A relationship-market model reduces the question to one quantity, the product of the curvature of companionship utility and the elasticity of substitution between human and machine company: above one, better companions crowd out human relationships; below one, they support them; either way they raise measured utility, so improvements can lift felt well-being while lowering the production of real relationships. A second result shows that reliance on companions, and exposure to their withdrawal, is concentrated among those with the poorest human-match prospects. The design uses the abrupt February 2023 removal of companionship features from a major application and the staggered arrival of policy restrictions across jurisdictions, with falsification tests that separate a companionship effect from the general rise of artificial intelligence. A Monte Carlo study, labeled as simulation throughout, confirms that the design recovers an imposed treatment path without bias, has adequate power at plausible effect sizes, and requires clustered inference under serial correlation. The post-shock data to execute it arrive over the next two to three years.

Keywords: artificial intelligence, loneliness, relationship markets, welfare, difference-in-differences, research design.

How selfish soever man may be supposed, there are evidently some principles in his nature, which interest him in the fortune of others, and render their happiness necessary to him, though he derives nothing from it except the pleasure of seeing it.

Adam Smith, *The Theory of Moral Sentiments* (1759)

1 Introduction

Loneliness in the United States is widespread and rising, and for the first time a large number of people are forming their closest daily relationships with software. Over roughly two years, applications that offer an artificial friend, confidant, or romantic partner went from a curiosity to a mass-market good used by tens of millions of people, many of whom report that the artificial companion is the voice they speak to most. The arrival of this good changes the technology of companionship itself, a good that is among the most consequential a person consumes. The question this paper poses is simple to state and hard to answer. When a cheap, patient, always-available artificial companion enters the market for human connection, does it substitute for the real thing or help people find it, and is the world better or worse for the trade.

The question matters because the stakes are large and because the policy clock is already running. The United States Surgeon General issued a formal advisory in May 2023 describing social isolation as an epidemic with health consequences on the order of smoking, a magnitude that rests on a large meta-analytic literature linking isolation to mortality [Holt-Lunstad et al., 2015]. In the same period, a regulator in Italy restricted a leading companionship application, a wrongful-death suit in October 2024 alleged that an artificial companion contributed to a teenager's suicide, and California enacted a statute in October 2025 requiring safety features in companion chatbots. Legislatures are writing rules for a good whose welfare effects no one has credibly measured. That is the ordinary situation for a new technology, and it is exactly the situation in which careful applied economics earns its

keep, provided the work is designed before it is run.

Two obstacles prevent a credible answer, and both shape the design. The first is that adoption is chosen, not assigned. The people who reach for an artificial companion are, on average, the people with the thinnest human options, so a raw comparison of users and non-users confounds the effect of the companion with the loneliness that drew them to it. The second is that artificial companions arrived inside a much larger wave. General-purpose generative models became a cultural event in November 2022, and any time series that moves with the spread of artificial companions also moves with the spread of artificial intelligence writ large. A credible design has to break both problems at once. It needs variation in artificial companionship that is plausibly unrelated to a person's underlying loneliness, and it needs that variation to be specific to companionship rather than to the broader technology.

The finding of this paper follows directly. There is no credible causal estimate of the effect of artificial companions on well-being or human relationships, and there cannot be one yet. The claims now in circulation, that companions relieve loneliness or that they deepen it, rest either on comparisons between users and non-users, who differ in the loneliness that led them to adopt, or on aggregate trends that rose alongside generative AI as a whole. Neither isolates the effect of the companion. The data that would isolate it, enough post-shock observations to estimate the design set out below, are only now accumulating. What the paper provides is the apparatus a credible estimate will require: a model that identifies what the answer depends on and reduces it to a single quantity, an identification strategy that locates that quantity, a confounder analysis that separates artificial companionship from the general rise of generative AI, falsification tests that distinguish a real effect from an artifact, and a demonstration that the design has adequate power to detect an effect of the size the theory implies. The approach follows a growing body of work on the welfare effects of digital technologies, including social media [Allcott et al., 2020, Braghieri et al., 2022] and recent descriptive evidence on artificial companions and loneliness [Maples et al., 2024]. The design follows the credibility tradition in applied microeconomics: identification rests on the

structure of the comparison, and the decisions that matter are specified before estimation.

The theory is a relationship-market model. A person divides scarce effort between human companionship, which is valuable but costly and uncertain to obtain, and machine companionship, which is cheap and reliable but carries an authenticity discount. The two combine into a single companionship aggregate, and the person maximizes the utility of that aggregate net of the cost of forming human ties and the price of the machine. The model delivers three lessons that organize the empirical work. First, whether better artificial companions crowd human connection out or draw it in turns on a single number, the product of the curvature of companionship utility and the elasticity of substitution between the two kinds of company. The sign of that product minus one is the object the design is built to recover. Second, improvements in artificial companionship raise measured utility whatever the sign of the crowd-out, so a policy that reads welfare off felt satisfaction can approve a change that is quietly reducing the stock of real human relationships. This is the wedge between measured utility and produced value, and it is the paper's central welfare result. Third, the people who rely most on artificial companions, and who therefore stand to lose the most when one is withdrawn or degraded, are precisely those with the poorest human-match prospects. That prediction gives the design a sharp test, because it says where the effect of a companionship shock should be concentrated.

The design turns on a real event. In February 2023, a leading companionship application abruptly removed the intimate features at the heart of its product, and an existing community of users experienced the change as the sudden loss of a relationship. The episode is close to an exogenous degradation of artificial companionship for people who had already selected into it, which is the variation identification requires. Around that event I build an event-study and difference-in-differences design that compares places and people with greater prior exposure to the application against those with less, before and after the shock, using search behavior, community text, help-seeking, and, over a longer horizon, mental-health and relationship outcomes. A second design uses the staggered arrival of policy restrictions

across jurisdictions, estimated with modern tools that are robust to treatment-effect heterogeneity, and a synthetic-control comparison for the single largest policy change. The main threat to this design is the general diffusion of generative artificial intelligence, which began at the same time as the companion applications and moves many of the same outcomes. The design addresses it through its structure. The February 2023 event is sharp and prior exposure to the affected application varies across places, so period fixed effects absorb any common national trend, and the broad technology would have to affect high-exposure places differentially at the exact date of a companionship-specific change in order to be mistaken for an effect. Placebo outcomes that respond to generative artificial intelligence in general but not to companionship provide a direct test. If the two cannot be separated, that is itself the result.

The remainder proceeds as follows. Section 2 describes the setting and the policy moment. Section 3 develops the model, states and proves the two central results, and draws out the welfare wedge and its policy implications. Section 4 describes the data. Section 5 lays out the identification strategy and its assumptions. Section 6 presents the falsification plan and confronts the confounder of record. Section 7 reports the simulation evidence that the design has power, labeled as simulation throughout. Section 8 discusses what a credible answer would mean for welfare and policy, and Section 9 concludes. Two appendices collect the full derivations and proofs.

2 The setting: a new good in the market for connection

Companionship has always been produced the hard way. People invest time, attention, and emotional risk to find friends and partners, the search is uncertain, and the market that clears these matches is thick for some and thin for others. The thinness is unequally distributed. It falls hardest on people whose circumstances, geography, age, health, or temperament leave them with few feasible partners, and for those people the cost of an additional unit of human

connection is high. The economics of this market has a long tradition, from the theory of marriage [Becker, 1973] and search models of matching by type [Burdett and Coles, 1997] to the empirical study of how technology changes who meets whom [Hitsch et al., 2010, Rosenfeld et al., 2019]. Into this market has entered a good that is engineered to be the opposite of the human one. An artificial companion is immediately available, infinitely patient, free or nearly free, and tuned to affirm. It never leaves, never judges, and never has a competing claim on its time. Whatever else it is, it is a dramatic reduction in the price and risk of a certain kind of companionship, and basic economics says that a large change in the price of a close substitute for a central consumption good will change behavior.

The scale is already significant. Leading companionship applications report user bases in the millions and engagement that rivals or exceeds that of mainstream social platforms, with sessions measured in appreciable fractions of the waking day for the most involved users. The population of users is not a uniform slice of the public. It tilts toward the isolated, the socially anxious, the bereaved, and the young, which is to say toward the thin side of the relationship market. That selection is the central obstacle to measurement, and it is also a clue to where the welfare consequences will land.

The policy environment has moved faster than the evidence. A European regulator restricted a major application in February 2023 on data-protection and minor-safety grounds. The United States Surgeon General's May 2023 advisory placed the loneliness epidemic on the national health agenda and quantified its mortality burden. A wrongful-death action in October 2024 put the question of whether an artificial companion can harm a vulnerable user into open court. California's October 2025 statute began to write the rules of the road for companion chatbots, and other jurisdictions are following. Each of these is a lever that moves the availability or the quality of artificial companionship for some people and not others, at a known moment, which is the raw material of a research design. The task of the next sections is to turn that raw material into a credible comparison, and to say in advance how we will know whether the comparison has worked.

3 A model of synthetic and human companionship

3.1 Environment and the agent's problem

A person chooses human companionship $h \geq 0$ and machine companionship $m \geq 0$. The two combine into a companionship aggregate through a constant-elasticity technology,

$$C(h, m) = \left[\alpha \, h^{\rho} + (1 - \alpha) \, (\phi m)^{\rho} \right]^{1/\rho}, \qquad \rho \leq 1, \ \rho \neq 0, \tag{1}$$

with taste weight $\alpha \in (0, 1)$ and elasticity of substitution $\sigma = 1/(1 - \rho)$. The parameter $\phi \in (0, 1]$ is the *authenticity* of machine companionship: one unit of machine time delivers only ϕ authenticity-equivalent units relative to a human unit. Preferences over the aggregate are $v(C)$ with $v' > 0$ and $v'' < 0$. Human companionship is costly to form and sustain, with cost $c(h; \theta)$ satisfying $c_h > 0$, $c_{hh} > 0$, and $c_{h\theta} < 0$, where θ indexes relationship-market prospects: a higher θ means a thicker market and a lower marginal cost of human connection. Machine companionship costs p per unit. The person solves

$$\max_{h, m \geq 0} \ \Pi(h, m; \phi, p, \theta) \ = \ v\big(C(h, m)\big) \ - \ c(h; \theta) \ - \ p \, m. \tag{2}$$

The authenticity ϕ and the price p are set by platforms and by policy. The February 2023 removal of companionship features is a discrete fall in ϕ; a ban is $p \to \infty$. Let $\gamma(C) \equiv -v''(C) \, C / v'(C)$ denote the curvature of companionship utility. Society values the produced stock of human relationships h beyond its private return, because human ties carry external benefits $e > 0$ through care, fertility, and social capital; the social objective is $\Pi + e \, h$. The gap between the private and social objectives drives the welfare analysis.

3.2 First-order conditions and the solution

The marginal products of the aggregate are $C_h = \alpha C^{1-\rho} h^{\rho-1}$, $C_m = (1-\alpha)\phi^\rho C^{1-\rho} m^{\rho-1}$, and $C_\phi = (1-\alpha) C^{1-\rho} \phi^{\rho-1} m^\rho > 0$. Interior optimality requires

$$v'(C)\, C_h = c_h(h;\theta), \tag{3}$$

$$v'(C)\, C_m = p. \tag{4}$$

Both conditions depend on (h, m) and the parameters only through the scalar $\mu \equiv v'(C)\, C^{1-\rho}$. From (4), $m^\star = \left[(1-\alpha)\phi^\rho \mu / p\right]^{1/(1-\rho)}$, which is increasing in μ. From (3), since c is convex and $\rho - 1 < 0$, the left side falls and the right side rises in h, so $h^\star(\mu, \theta)$ is unique with $\partial h^\star / \partial \mu > 0$ and $\partial h^\star / \partial \theta > 0$. The objective is strictly concave on the interior when v is concave, c is convex, and the Hessian is negative definite, which holds for sufficiently convex human-connection cost and is verified numerically at the calibrated optimum in Appendix B.

Assumption 1 (Regularity)**.** An interior optimum exists with negative-definite Hessian and $\partial C^\star / \partial \phi > 0$.

3.3 The pivotal quantity: crowd-out versus crowd-in

The central comparative static concerns how better artificial companionship moves human connection.

Theorem 1 (Crowd-out and the utils-versus-production wedge)**.** *Under Assumption 1, at the interior optimum:*

(a) Private experienced utility is strictly increasing in authenticity, $\partial \Pi^\star / \partial \phi = v'(C^\star)\, C_\phi > 0$.

(b) The effect on human companionship is governed by a single sufficient statistic,

$$\operatorname{sign}\left(\frac{\partial h^\star}{\partial \phi}\right) = \operatorname{sign}\big(\,1 - \gamma(C^\star)\,\sigma\,\big). \tag{5}$$

Human companionship falls when artificial companionship improves (crowd-out) if and only if $\gamma\sigma > 1$; it rises (crowd-in) if and only if $\gamma\sigma < 1$.

Consequently, whenever $\gamma\sigma > 1$, an improvement in artificial companionship raises measured utility while reducing the production of human relationships.

The proof is in Appendix A; the logic is short. Part (a) is the envelope theorem: at the optimum only the direct effect of ϕ survives, and it is positive. Part (b) follows because $h^\star$ depends on ϕ only through μ, so the sign of $\partial h^\star/\partial\phi$ equals the sign of $\partial\mu^\star/\partial\phi$, and differentiating $\mu = v'(C)C^{1-\rho}$ gives $d\mu/dC = C^{-\rho}v'\big[(1-\rho)-\gamma\big]$, whose sign is that of $1/\sigma - \gamma$, equivalently $1-\gamma\sigma$. The economics is that two forces pull in opposite directions when artificial company improves. Substitution pushes effort away from the now relatively expensive human option, while the fall in the marginal value of the companionship aggregate, through the curvature of v, changes how much total companionship the person wants. The product $\gamma\sigma$ is exactly the rate at which these forces trade off. The figure makes the two regimes visible: Figure 2 shows human connection falling as authenticity rises when $\gamma\sigma > 1$ and rising when $\gamma\sigma < 1$, and Figure 3 shows experienced utility climbing in both cases.

The wedge is the central welfare result. A planner who measures welfare by elicited satisfaction records an improvement whenever ϕ rises, because felt utility does rise. In the region where human and artificial company are gross substitutes, however, the same improvement reduces the stock of human relationships, the produced good whose external value e the individual does not internalize. Measured utility and produced value move in opposite directions, so a welfare standard based on elicited satisfaction misstates the effect of the change.

3.4 Who bears the shock

The second result concerns incidence, and it gives the design its sharpest prediction.

Theorem 2 (Incidence: the lonely bear the shock)**.** *Under Assumption 1, in the substitutes regime $\gamma\sigma > 1$:*

(i) Reliance on artificial companionship falls in relationship-market prospects, $\partial m^\star/\partial\theta < 0$: the lonely use artificial companions more.

(ii) The marginal value of authenticity falls in prospects, $\partial^2\Pi^\star/\partial\phi\,\partial\theta < 0$, so a negative authenticity shock imposes a larger welfare loss on low-θ agents.

The proof is in Appendix A. The intuition is that better human prospects raise human companionship and the companionship aggregate, which in the substitutes regime lowers the shadow value μ and hence reduces machine use; and because the welfare value of authenticity runs through machine use, those who use the machine most are those who lose most when it is taken away. The empirical content is immediate and testable. The effect of the February 2023 shock should not be uniform. It should be concentrated in thin relationship markets, among people identifiable by low local partner availability, rurality, age, and prior isolation. A design that finds a diffuse, undifferentiated response would be failing to confirm the mechanism the theory predicts.

Corollary 1 (Policy)**.** *A planner who values $\Pi + e\,h$ internalizes the relational externality. In the substitutes regime the private optimum over-consumes artificial companionship relative to the social optimum, and a quality floor or an age restriction that raises the effective human share is corrective. In the complements regime the same restriction reduces welfare. The sign of $\gamma\sigma - 1$ is therefore the pivotal quantity for policy as well as for science, which is why recovering it is the object of the design.*

4 Data

The empirical program draws on nine sources, summarized in Table 1, chosen so that the central phenomena can be triangulated rather than rested on any single series. The design treats some sources as measures of exposure, some as outcomes, and some as the raw material of the treatment timing itself.

Exposure to artificial companionship is measured two ways. Search behavior from a public search- index provides companion-application interest at the level of media markets and states by week, back to the mid-2000s, which supplies both a pre-shock exposure measure and a high-frequency outcome. Commercial application-analytics data provide downloads and session time at the application-month level, a more direct measure of engagement that can validate the search-based proxy. Community text archived from the platforms themselves gives a window into the mechanism, because the reaction to the February 2023 shock was written down in real time by the people who experienced it.

The outcomes run from proximate to distal. The most proximate are high-frequency behaviors, search for distress and help, and the volume of contacts to crisis lines, available by state and month for the recent period. Intermediate outcomes come from repeated health surveys that measure psychological distress and isolation at the state-year level, and from the national time-use survey, which records time spent alone and time spent with friends and can serve as both an outcome and a mechanism. The most distal outcome, and the one that matters most for policy, is mortality from suicide and self-harm, available by county and month over a long horizon, which the design can only address once enough post-period data have accumulated. Relationship-market outcomes, how couples meet and the prevalence of online dating, come from a long-running survey of American couples and from national survey series. Finally, the treatment timing itself is read from the public record of regulatory actions and statutes across jurisdictions, which supplies the staggered variation used in the second design.

5 Research design and identification

5.1 The estimand

The target is a treatment effect of the authenticity and availability of artificial companionship on human well-being and human relationship formation, defined over a named population and with an explicit weighting. In the event-study design the estimand is the dynamic average effect, per unit of prior exposure, of the February 2023 degradation of companionship features, which under the model recovers the sign of $\gamma\sigma - 1$. In the policy design the estimand is a collection of group-time average treatment effects on the treated, one for each jurisdiction and time since restriction, aggregated with transparent weights. Both estimands are stated in potential-outcome terms so that the identifying assumptions attach to objects the design can be built around.

5.2 Design one: the authenticity shock

The first design exploits the abrupt February 2023 removal of companionship features. Let Y_{it} be an outcome in region i and period t, let D_i be prior exposure to the affected application measured in the pre-period, and let the shock occur at t_0. The event-study specification is

$$Y_{it} = \sum_{k \neq -1} \beta_k \mathbb{1}[t = t_0 + k] \times D_i + \alpha_i + \delta_t + \varepsilon_{it}, \tag{6}$$

with region and period fixed effects and inference clustered by region. The coefficients β_k for $k \geq 0$ trace the dynamic effect of losing companionship quality in high-exposure places, and the coefficients for $k < 0$ are the pre-trend test. Identification requires parallel trends, that high- and low-exposure regions would have moved together absent the shock, and no-anticipation, that behavior did not adjust before the unannounced change. The appeal of the event is that the degradation fell on an existing user base for reasons internal to the platform, so it is plausibly unrelated to the contemporaneous evolution of loneliness across

places.

5.3 Design two: staggered policy restrictions

The second design uses the staggered arrival of restrictions on artificial companions across jurisdictions. Because treatment timing varies and effects are likely heterogeneous, the two-way fixed-effects estimator is biased [de Chaisemartin and D'Haultfœuille, 2020, Goodman-Bacon, 2021], and the design instead uses group-time average treatment effects on the treated with not-yet-treated comparisons [Callaway and Sant'Anna, 2021], an interaction-weighted event study [Sun and Abraham, 2021], and a decomposition that makes the weights transparent [Goodman-Bacon, 2021]. For the single largest policy change the design adds a synthetic-control comparison [Abadie et al., 2010, Abadie, 2021], constructing a weighted combination of untreated jurisdictions that reproduces the treated unit's pre-period path and reading the effect off the post-period gap, with inference by permutation. The two designs answer the same question through different variation, which is the triangulation the program is built to deliver.

6 Falsification and the confounder of record

A design is only as credible as the alternatives it can rule out, and the alternatives here are specific enough to be confronted one by one. Table 4 maps each threat to the test that addresses it.

The confounder of record is the general salience of generative artificial intelligence. Artificial companions diffused inside the broader wave that began in November 2022, and any series that rises with companion adoption also rises with the spread of the underlying technology. The design separates the two with placebo outcomes. A companionship-specific shock, the removal of intimate features from one application, should move companionship-related behavior but should not move search or engagement tied to general-purpose artificial

intelligence that was unaffected by the change. If the placebo outcome moves with the shock, the companionship interpretation is in doubt and the paper says so. If it does not, the companionship-specific channel is isolated. The same logic disciplines the second design, where only companion-specific policies, not general technology policy, should bite.

The remaining threats are handled in the standard way, pre-specified rather than discovered after the fact. Differential pre-trends are tested directly in the event-study pre-period and bounded with a sensitivity analysis that asks how large a trend violation would have to be to overturn the result [Rambachan and Roth, 2023]. Reverse causality, that distress drives adoption rather than the other way around, is addressed by the timing and the exogeneity of a platform-initiated change to the individual user. Coincidental timing is tested with placebo shock dates on which no policy or platform change occurred, and spurious exposure with placebo regions that had essentially no pre-shock exposure. Influential units and selection are probed with leave-one-region-out re-estimation and permutation inference. Most important, the mechanism itself is a falsification: Theorem 2 predicts that the response is concentrated in thin relationship markets, so a finding that the effect is uniform across markets would be evidence against the channel, not merely a null on a subgroup.

If the companionship-specific variation cannot be cleanly separated from the general-technology wave, the paper characterizes the confounder directly. An account of how the salience of artificial intelligence as a whole co-moves with loneliness and help-seeking is itself informative, and it is the result the data would support.

7 Design evaluation: what the design can detect

Before any real outcome is examined, a design should be made to prove that it can recover a known answer and that it has the power to see an effect of the size the theory implies. This section reports that evaluation. Everything in it is simulation. No real data are used, and no number here is an estimate of anything in the world. The point of the exercise is the

design, not a finding.

The simulation builds a balanced panel of regions and periods, assigns each region a continuous prior-exposure value, imposes a known dynamic treatment path that is zero before the shock and ramps to a plateau after it, and adds region and period fixed effects together with serially correlated idiosyncratic noise. It then estimates the event-study specification in (6) and the single-coefficient difference-in-differences that underlies the power analysis. Three results follow. First, Figure 4 shows that the estimated dynamic path sits on top of the imposed path, with pre-period coefficients indistinguishable from zero, so the design is unbiased when its assumptions hold. Second, Figure 5 traces statistical power against effect size for panels of different widths, and Table 3 reports performance at a benchmark effect, showing that a design with on the order of a hundred regions detects plausibly sized effects with high probability while a much smaller design does not, which is a quantitative statement about how much data the real study will need. Third, Figure 6 and Table 5 make the inference point concrete: when the idiosyncratic errors are serially correlated, as outcomes like search and mortality surely are, naive standard errors reject a true null far too often, and only clustered inference restores the nominal size. The lesson is the one that difference-in-differences practice learned the hard way [Bertrand et al., 2004], and the design incorporates it from the start.

8 Discussion: welfare, policy, and the value of connection

Suppose the design is executed and the sign of $\gamma\sigma - 1$ is recovered. What would each outcome mean. If artificial and human companionship are gross substitutes, the policy question is no longer whether artificial companions make their users feel better, because the model says they do, and so, most likely, will the data. The question is whether the felt improvement is bought at the cost of the slow erosion of the human relationships that carry

external value, the friendships and partnerships and families whose benefits spill beyond the individual. In that world a regulator who certifies artificial companions on the basis of user satisfaction is measuring the wrong thing, and the corrective instruments are the ones legislatures are already reaching for, quality floors and age restrictions that preserve a minimum human share, now with a reason behind them. If instead the two are complements, artificial companions are scaffolding that helps isolated people re-enter the human market, and restricting them would remove a ladder from the people who most need it. The same statute is good policy in one regime and bad policy in the other, which is why the sign is worth the cost of measuring carefully.

The incidence result sharpens the stakes. Because reliance on artificial companions is concentrated among those with the poorest human options, both the benefits and the harms of this technology fall disproportionately on the already lonely. A good that raises the felt well-being of the isolated while loosening their remaining grip on human connection is a distributionally loaded object, and the usual summary statistics will hide that. The right evaluation has to look at the tail, not the average, which is where the model says the action is and where the design is built to look.

There is a longer view. The closest precedents are the welfare studies of social media [Allcott et al., 2020, Braghieri et al., 2022] and the debate over screens and adolescent mental health [Twenge et al., 2018], which found that a technology can be privately valued and socially costly at once. Standard welfare analysis treats well-being as elicited satisfaction, and for most goods that convention is adequate. It is not adequate here, because an artificial companion can raise felt connection while reducing the human relationships that produce it. The wedge between measured utility and produced value, usually an abstraction, is first-order for companionship, which is why the question warrants a credible research design.

9 Conclusion

Artificial companions are a large and growing presence in private life, they are being regulated before they are understood, and the users most exposed to them are those least able to absorb a mistake. The finding of this paper is that their effect on well-being and on human relationships is not yet known, and that the claims to the contrary are not credible. Those claims come in two forms. The first compares people who use artificial companions with people who do not, and attributes the difference to the companion; but users select into adoption on the basis of the loneliness that is itself the outcome of interest, so the comparison recovers selection, not an effect. The second reads a causal story off aggregate time trends; but artificial companions diffused together with generative artificial intelligence as a whole, and the two cannot be told apart in a national time series. Until a design separates the companion from the general technology, and until enough post-shock data exist to estimate it, the honest answer to whether artificial companions help or harm is that the evidence does not yet support either conclusion.

The negative finding does not imply inaction. The paper identifies what the answer depends on, a single pivotal quantity, the product of the curvature of companionship utility and the elasticity of substitution between human and machine company. It builds an identification strategy around a real degradation of authenticity and a body of staggered policy variation, specifies the falsifications that separate a companionship effect from the broader diffusion of artificial intelligence, and shows by simulation that the design has the power to detect an effect of the magnitude the theory implies. The next two to three years will supply the post-shock data the design requires. The apparatus to use them is set out here.

References

Alberto Abadie. Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425, 2021. doi: 10.1257/jel.20191450.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505, 2010. doi: 10.1198/jasa.2009.ap08746.

Hunt Allcott, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. The welfare effects of social media. *American Economic Review*, 110(3):629–676, 2020. doi: 10.1257/aer.20190658.

Gary S. Becker. A theory of marriage: Part i. *Journal of Political Economy*, 81(4):813–846, 1973. doi: 10.1086/260084.

Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, 119(1):249–275, 2004. doi: 10.1162/003355304772839588.

Luca Braghieri, Ro'ee Levy, and Alexey Makarin. Social media and mental health. *American Economic Review*, 112(11):3660–3693, 2022. doi: 10.1257/aer.20211218.

Ken Burdett and Melvyn G. Coles. Marriage and class. *The Quarterly Journal of Economics*, 112(1):141–168, 1997. doi: 10.1162/003355397555154.

Brantly Callaway and Pedro H. C. Sant'Anna. Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230, 2021. doi: 10.1016/j.jeconom.2020.12.001.

Clément de Chaisemartin and Xavier D'Haultfœuille. Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–2996, 2020. doi: 10.1257/aer.20181169.

Andrew Goodman-Bacon. Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277, 2021. doi: 10.1016/j.jeconom.2021.03.014.

Günter J. Hitsch, Ali Hortaçsu, and Dan Ariely. Matching and sorting in online dating. *American Economic Review*, 100(1):130–163, 2010. doi: 10.1257/aer.100.1.130.

Julianne Holt-Lunstad, Timothy B. Smith, Mark Baker, Tyler Harris, and David Stephenson. Loneliness and social isolation as risk factors for mortality: A meta-analytic review. *Perspectives on Psychological Science*, 10(2):227–237, 2015. doi: 10.1177/1745691614568352.

Bethanie Maples, Merve Cerit, Aditya Vishwanath, and Roy Pea. Loneliness and suicide mitigation for students using GPT3-enabled chatbots. *npj Mental Health Research*, 3(1):4, 2024. doi: 10.1038/s44184-023-00047-6.

Ashesh Rambachan and Jonathan Roth. A more credible approach to parallel trends. *The Review of Economic Studies*, 90(5):2555–2591, 2023. doi: 10.1093/restud/rdad018.

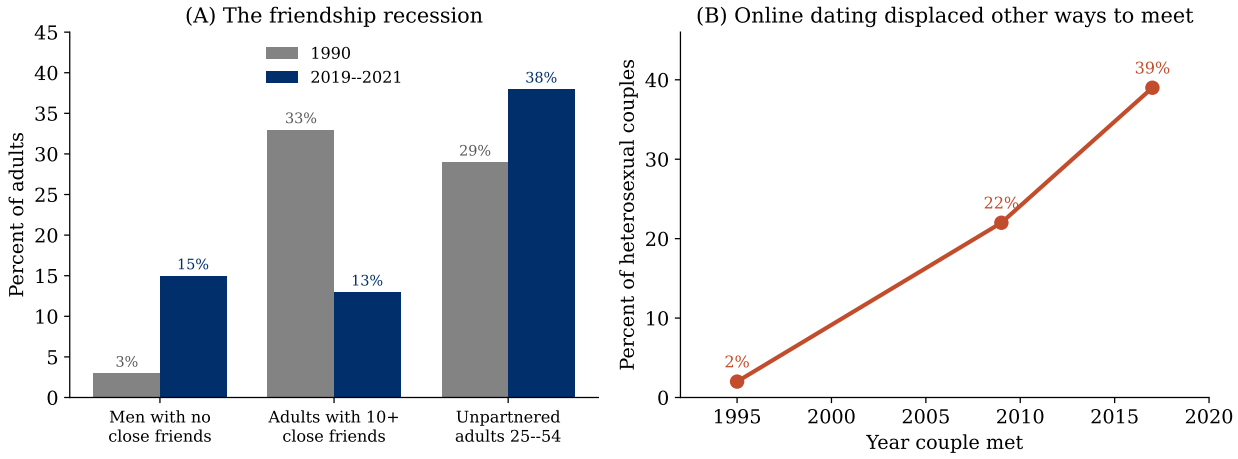
Michael J. Rosenfeld, Reuben J. Thomas, and Sonia Hausen. Disintermediating your friends: How online dating in the united states displaces other ways of meeting. *Proceedings of the National Academy of Sciences*, 116(36):17753–17758, 2019. doi: 10.1073/pnas.1908630116.

Liyang Sun and Sarah Abraham. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199, 2021. doi: 10.1016/j.jeconom.2020.09.006.

Jean M. Twenge, Thomas E. Joiner, Megan L. Rogers, and Gabrielle N. Martin. Increases in depressive symptoms, suicide-related outcomes, and suicide rates among u.s. adolescents after 2010 and links to increased new media screen time. *Clinical Psychological Science*, 6 (1):3–17, 2018. doi: 10.1177/2167702617723376.

Figures

Figure 1. Human connection has thinned while technology remade the relationship market



Sources: Survey Center on American Life (Cox 2021); Pew Research / U.S. Census; Rosenfeld, Thomas and Hausen (2019). Descriptive series, not estimates.

Figure 1: The two trends that frame the question: the long decline of human connection and the rapid arrival of artificial companionship. Descriptive series from the sources noted; not estimates.

Figure 2. Model-implied response of companionship to AI authenticity

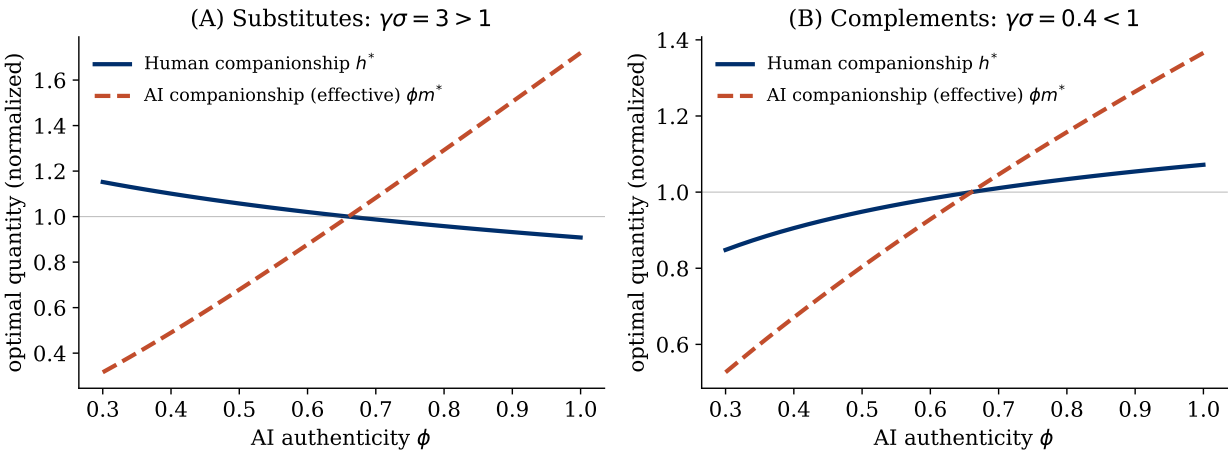


Figure 2: Model-implied response of human and artificial companionship to authenticity ϕ. In the substitutes regime ($\gamma\sigma > 1$) human connection falls as artificial companionship improves; in the complements regime ($\gamma\sigma < 1$) it rises. Model-implied, not estimated.

Figure 3. Welfare wedge and the incidence of an AI-companion shock

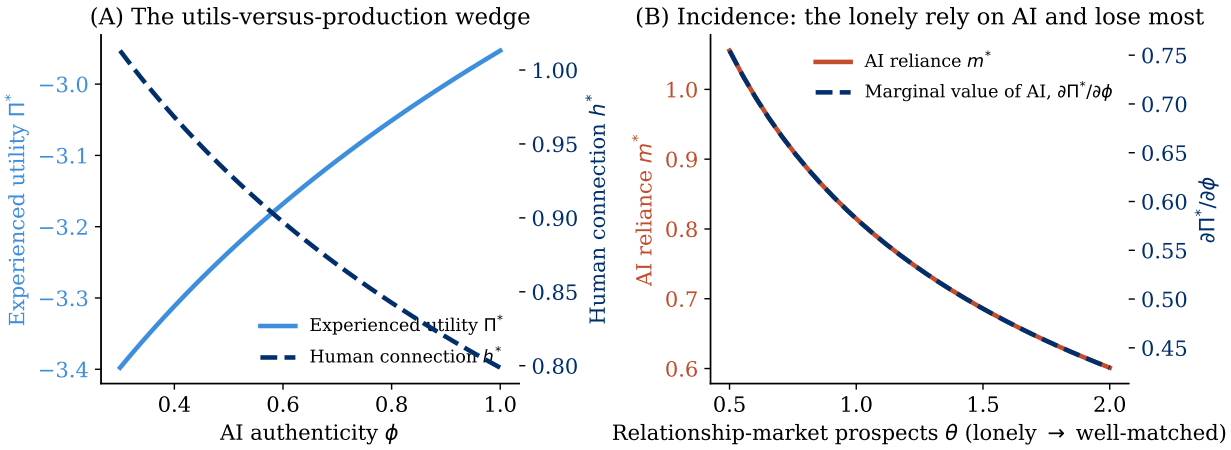


Figure 3: Left: experienced utility rises in authenticity while human connection falls, the utils-versus-production wedge. Right: reliance on artificial companionship and the marginal value of authenticity both fall in relationship-market prospects, so the lonely lose most (Theorem 2). Model-implied, not estimated.

Figure 4. SIMULATION: the design recovers the imposed dynamic path

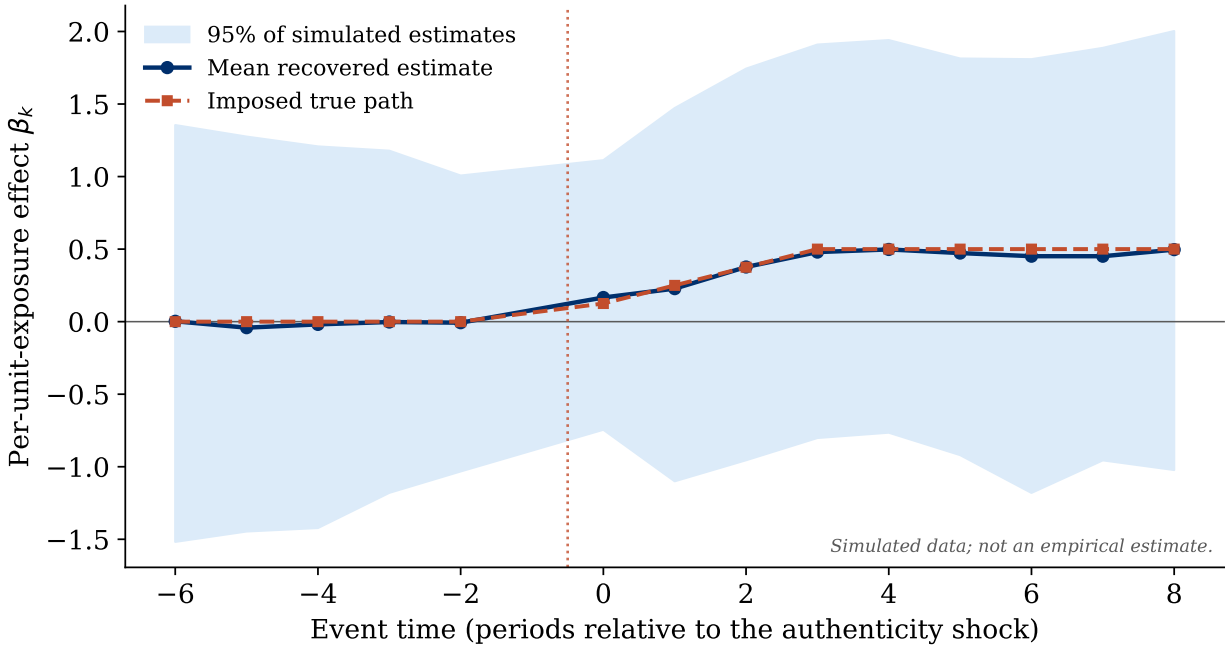


Figure 4: SIMULATION. The event-study design recovers an imposed dynamic treatment path with pre-period coefficients at zero. Simulated data; not an empirical estimate.

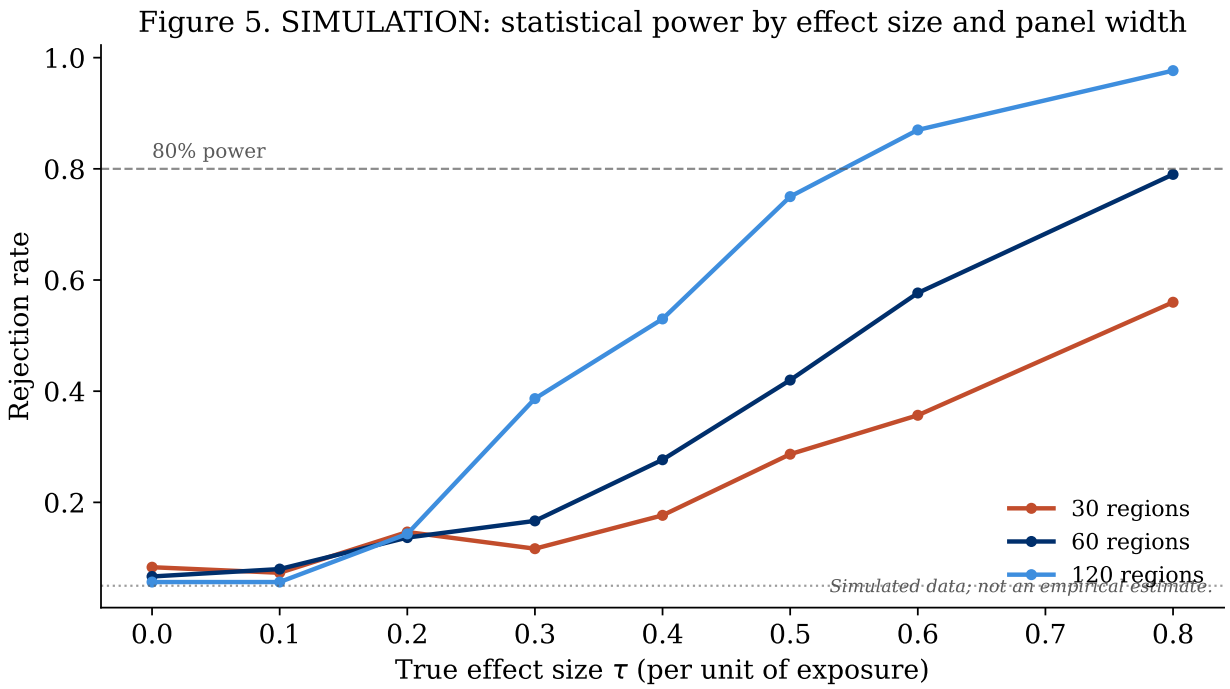


Figure 5: SIMULATION. Statistical power against effect size for panels of different widths. Simulated data; not an empirical estimate.

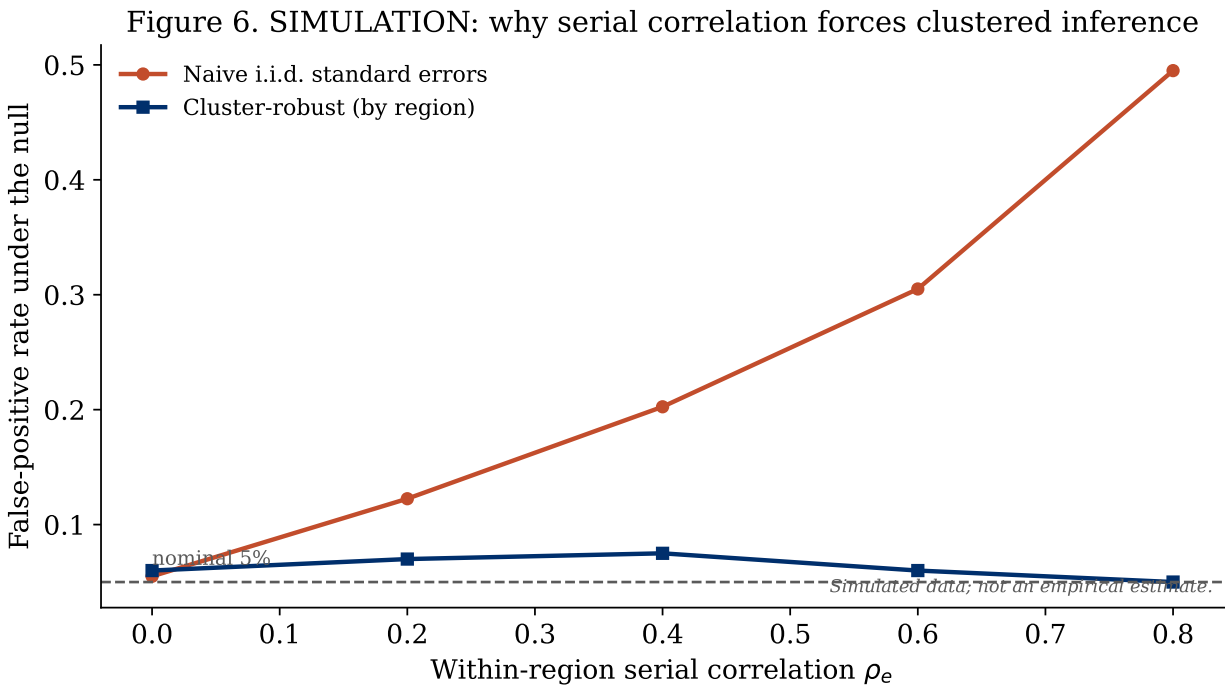


Figure 6: SIMULATION. With serially correlated errors, naive standard errors over-reject a true null while clustered inference holds nominal size. Simulated data; not an empirical estimate.

Tables

Table 1: Data sources and their role in the design.

Source	Variable (role in design)	Unit / coverage	Access
Google Trends	Companion-app and distress search intensity (exposure and outcome proxy)	DMA/state × week, 2004–present	Public API
Archived platform text (r/Replika, r/CharacterAI)	Community sentiment around the Feb. 2023 shock (mechanism)	Post-level	Public archive
Sensor Tower / data.ai	App downloads and session time (exposure)	App × month	Commercial
CDC WONDER	Suicide and self-harm mortality (downstream outcome)	County × month, 1999–present	Public
NSDUH / BRFSS	Serious psychological distress, isolation (outcome)	State × year	Public
American Time Use Survey	Time spent alone and socializing (outcome, mechanism)	Person-day, 2003–present	Public
988 / crisis-line volumes	Help-seeking (outcome)	State × month, 2022–present	Public / FOIA
HCMST (Rosenfeld) / Pew	How couples meet; online dating use (relationship market)	Repeated survey	Public
Policy records (Garante, SB 243, state age-gates)	Treatment timing (staggered DiD)	Jurisdiction × date	Legal records

Table 2: Model-implied comparative statics by regime, with the pivotal statistic $\gamma\sigma$.

	Substitutes ($\gamma\sigma > 1$)	Complements ($\gamma\sigma < 1$)	Pivotal statistic
$\gamma\sigma$ (RRA $\times$ elast. subst.)	3.00	0.40	threshold $= 1$
$\partial m^*/\partial\phi$ (AI own effect)	$+$	$+$	—
$\partial h^*/\partial\phi$ (crowd-out)	$-$	$+$	sign$(1-\gamma\sigma)$
$\partial\Pi^*/\partial\phi$ (private welfare)	$+$	$+$	always $+$
$\partial m^*/\partial\theta$ (lonely use AI more)	$-$	—	Thm. 2(i)
$\partial^2\Pi^*/\partial\phi\,\partial\theta$ (incidence)	$-$	—	Thm. 2(ii)
Hessian negative definite	yes	yes	SOC

Table 3: SIMULATION. Design performance: bias, interval coverage, and power by panel width.

Design performance (SIMULATION)	30 regions	60 regions	120 regions
Mean post-period bias in β_k		-0.015	
95% interval coverage of truth		1.00	
Power at $\tau = 0.5$	0.29	0.42	0.75
Power at $\tau = 0.3$	0.12	0.17	0.39

Table 4: Falsification plan: each threat and the test that addresses it.

Threat	Test	What it would rule out
General generative-AI salience (ChatGPT, Nov. 2022)	Placebo outcome: general-AI search unrelated to companionship should not move at the companion-specific shock	Separates companion-AI from general-AI hype (the confounder of record)
Differential pre-trends	Event-study pre-period coefficients ≈ 0; Rambachan–Roth sensitivity bounds	Nonparallel trends
Reverse causality (distress drives adoption)	Timing: the platform shock is exogenous to the individual; test leads vs. lags	Simultaneity
Coincidental timing	Placebo shock dates with no policy or platform change	Spurious common shocks
Spurious exposure	Placebo regions with near-zero pre-shock exposure	Omitted region-level shocks
Composition / outliers	Balanced panel, leave-one-region-out, permutation inference	Selection and influential units
Mechanism (Theorem 2)	Heterogeneity by thin markets (sex ratio, rurality, age, prior isolation)	Confirms the substitution channel

Table 5: SIMULATION. Size under the null: naive versus clustered standard errors by serial correlation.

Serial correlation ρ_e	0.0	0.2	0.4	0.6	0.8
Naive i.i.d. SE (size)	0.055	0.122	0.203	0.305	0.495
Cluster-robust SE (size)	0.060	0.070	0.075	0.060	0.050

SIMULATION; nominal size = 0.05. Null $\tau = 0$.

Table 6: Policy variation available for the staggered design.

Event / jurisdiction	Date	Nature	Usable variation
Italy (Garante)	Feb. 2, 2023	Order barring Replika from processing Italian users' data (minor safety)	Synthetic control; interrupted time series
Replika feature removal	early Feb. 2023	Abrupt removal of companionship features for an existing user base	Event study (authenticity shock)
Garcia v. Character Techs.	Oct. 22, 2024	Wrongful-death suit; national salience shock	Interrupted time series
California SB 243	Signed Oct. 13, 2025; effective Jan. 1, 2026	Companion-chatbot disclosure, minor safeguards, self-harm protocols	Staggered DiD (state)
New York companion law	Effective Nov. 5, 2025	Disclosure and self-harm safeguards	Staggered DiD (state)

Appendix A: Proofs

Proof of Theorem 1. The objective is $\Pi(h, m; \phi) = v(C(h, m)) - c(h; \theta) - pm$ with C the constant-elasticity aggregate. By the envelope theorem, at the optimum $\partial \Pi^\star / \partial \phi = v'(C^\star)\, C_\phi$, and $C_\phi = (1-\alpha) C^{1-\rho} \phi^{\rho-1} m^\rho > 0$, which proves part (a). For part (b), note that both first-order conditions (3) and (4) depend on the choice variables and parameters only through $\mu \equiv v'(C) C^{1-\rho}$. Condition (3) reads $\alpha \mu h^{\rho-1} = c_h(h; \theta)$; the left side is strictly decreasing in h (since $\rho < 1$) and strictly increasing in μ, while the right side is strictly increasing in h (since c is convex). Hence $h^\star$ is a strictly increasing function of μ, so $\operatorname{sign}(\partial h^\star / \partial \phi) = \operatorname{sign}(\partial \mu^\star / \partial \phi)$. Differentiating $\mu = v'(C) C^{1-\rho}$,

$$\frac{d\mu}{dC} = v'' C^{1-\rho} + (1-\rho) v' C^{-\rho} = C^{-\rho} v' \big[(1-\rho) - \gamma(C) \big],$$

which has the sign of $(1-\rho) - \gamma = 1/\sigma - \gamma$, equivalently $1 - \gamma\sigma$. Under Assumption 1, $\partial C^\star / \partial \phi > 0$ and $C^{-\rho} v' > 0$, so $\operatorname{sign}(\partial \mu^\star / \partial \phi) = \operatorname{sign}(1 - \gamma\sigma)$. Combining, $\operatorname{sign}(\partial h^\star / \partial \phi) = \operatorname{sign}(1 - \gamma\sigma)$. The wedge statement follows from part (a) and part (b) jointly: when $\gamma\sigma > 1$, $\partial \Pi^\star / \partial \phi > 0$ while $\partial h^\star / \partial \phi < 0$. ■

Proof of Theorem 2. (i) From $m^\star = \big[(1-\alpha) \phi^\rho \mu / p\big]^{1/(1-\rho)}$, $\operatorname{sign}(\partial m^\star / \partial \theta) = \operatorname{sign}(\partial \mu^\star / \partial \theta)$. A higher θ lowers the marginal cost of human connection ($c_{h\theta} < 0$), raising $h^\star$ and hence $C^\star$, so $\partial C^\star / \partial \theta > 0$. With $d\mu / dC = C^{-\rho} v' [(1-\rho) - \gamma] < 0$ in the regime $\gamma\sigma > 1$, it follows that $\partial \mu^\star / \partial \theta < 0$ and therefore $\partial m^\star / \partial \theta < 0$. (ii) By the envelope theorem $\partial \Pi^\star / \partial \phi = v'(C^\star)(1-\alpha) C^{\star\, 1-\rho} \phi^{\rho-1} m^{\star\, \rho}$, which is increasing in $m^\star$. Since $\partial m^\star / \partial \theta < 0$ by part (i), the cross-partial $\partial^2 \Pi^\star / \partial \phi \, \partial \theta < 0$. A fall in ϕ therefore reduces $\Pi^\star$ by more for low-θ agents. ■

Appendix B: Comparative statics and numerical verification

This appendix collects the full comparative-statics algebra and reports the numerical verification that underlies the figures and Table 2. The functional forms used for the illustrations are constant relative risk aversion over the aggregate, $v(C) = C^{1-\gamma}/(1-\gamma)$, and a quadratic human-connection cost, $c(h;\theta) = h^2/(2\theta)$, so that $c_h = h/\theta$, $c_{hh} = 1/\theta > 0$, and $c_{h\theta} = -h/\theta^2 < 0$. With these forms γ is a constant and $\sigma = 1/(1-\rho)$ is a constant, so the pivotal product $\gamma\sigma$ is a single number that places the economy in the substitutes or the complements regime once and for all.

The sufficient-statistic reduction. The two first-order conditions $\alpha\mu h^{\rho-1} = h/\theta$ and $(1-\alpha)\phi^\rho \mu m^{\rho-1} = p$ determine $h^\star$ and $m^\star$ as functions of the scalar $\mu = v'(C)C^{1-\rho}$ and the parameters. Solving the first for h gives $h^\star = \big(\alpha\theta\mu\big)^{1/(2-\rho)}$, and the second gives $m^\star = \big[(1-\alpha)\phi^\rho\mu/p\big]^{1/(1-\rho)}$. Substituting both into the definition $C = [\alpha h^\rho + (1-\alpha)(\phi m)^\rho]^{1/\rho}$ yields a single fixed-point equation in μ, which has a unique solution under Assumption 1. All comparative statics in the main text are obtained by differentiating this fixed point and applying the sign of $d\mu/dC$ derived in Appendix A.

Numerical verification. The script `code/python/model.py` solves the model at the calibrated parameters for two regimes, a substitutes regime with $\sigma = 2$ and $\gamma = 1.5$ so that $\gamma\sigma = 3$, and a complements regime with $\sigma = 0.5$ and $\gamma = 0.8$ so that $\gamma\sigma = 0.4$. In both the Hessian of the objective is negative definite at the optimum, so the second-order conditions hold. In the substitutes regime the solver returns $\partial h^\star/\partial\phi < 0$, $\partial\Pi^\star/\partial\phi > 0$, $\partial m^\star/\partial\theta < 0$, and $\partial^2\Pi^\star/\partial\phi\,\partial\theta < 0$, confirming Theorems 1 and 2. In the complements regime $\partial h^\star/\partial\phi > 0$, confirming the sign reversal. The predicted and realized signs of $\partial h^\star/\partial\phi$ agree with $\text{sign}(1-\gamma\sigma)$ in both cases. These results are reproduced independently in R and Stata; the cross-language reconciliation is reported in `code/replication/`.