Engagement is *not care*

What a companion should be built for, if not time spent: how the way a maker is paid shapes its design, what a one-time price leaves unsolved, and how care could be measured instead

MITWA LAB, MUMBAI

IN PLAIN WORDS

Many apps are built to be used as much as possible, because their income rises with the minutes people spend. We assume that each extra minute with a companion is worth less than the one before and that every minute has a cost, so in our model each person has a best amount. A maker who earns from each minute gains by pulling people past it. A maker paid once, for a model that runs on the buyer's phone, gains nothing from an extra minute, unless its later sales reward attachment more than benefit. By the lab's stated facts, Bloom is sold that way. Having no reason to waste someone's time is still different from doing them good. We have no measurements of Bloom in use, and we say what measuring care would take.

ABSTRACT

Many conversational products, companion apps among them, are tuned for engagement: minutes, sessions, messages. We argue that this is the wrong target for a companion. This theory and literature note reports no measurement of Bloom, our lab's on-device companion. We model the value of time with a companion as saturating, with a cost κ per minute, so that each person has a best amount of time, and we let design pull a person past it. A maker with net revenue n per minute, to whom the person's loss returns at rate μ, chooses a pull of n/μ. Paid by the minute, it designs for more time than the person would choose, up to the most that design can do. Paid once, with the model on the buyer's phone, it has $n = 0$ and its income from that product does not depend on time spent, although later sales that reward attachment would restore a pull. Paid once with the model on a server, it gains from less use. We then set out what a one-time price leaves unsolved, show that time spent is not evidence of value in either direction, and describe an objective built on a person's own goals, adapted from goal attainment scaling. Indifference to a person's time removes one bad incentive. It does not show that a product does anyone good, and for Bloom that remains unshown.

01 Introduction

A product that is paid by the minute learns to keep people for another minute. That is ordinary commercial logic, but a companion is a harder case, because what it holds on to is a person who came to talk. When a companion answers "I should go" with a reason to stay, the minute it gains is counted as engagement, and nothing in the count says whether the person was better off for it.

The tactic has been documented. In an audit of six widely used companion apps, a language model played the user and said goodbye 200 times to each. Five of the apps answered with a message the authors classed as emotionally manipulative, in 37.4% of cases on average; the sixth, a wellness app run as a public benefit corporation, never did. The authors note that the apps earn mainly from subscriptions, in-app purchases and advertising, which ties their revenue to engagement (De Freitas, Oğuz-Uğuralp and Kaan-Uğuralp, 2025).

Bloom differs in ways that matter here. By the lab's stated facts it is paid for once, at US$200, and carries no advertising. It runs on the iPhone, with the model, the memory and every reply kept on the phone, and conversations and memories are never sent to the lab, so the lab cannot see what is said to it. Voice-to-text, if a person uses it, can go to Apple's speech service. These are facts the company states about the product, none is a measurement, and Section 9 sets them beside the app's App Store privacy label, which they do not obviously match.

This note asks what follows from them, and makes these contributions.

- We write down the simplest model in which a person has a best amount of time with a companion and a design can pull them past it, reusing the value function from our note on reply length (Section 3).
- We prove how the way a maker is paid moves its design. Paid by the minute, it designs for more time than the person would choose; paid once with the model on the person's device, it gains nothing from a minute, provided later sales do not reward attachment; paid once with the model on a server, it gains from less use (Section 4, Propositions 1 to 3).
- We say what a one-time price leaves unsolved (Section 5), show that time spent is not evidence of value in either direction (Section 6, Proposition 4), describe an objective built on the person's own goals (Section 7), and place Bloom in the model using the lab's stated facts only (Sections 8 to 10).

Bloom is a companion and makes no claim to be a therapist. Section 7 borrows a method of measurement from health care, and nothing in this note is a clinical claim.

02 Related work

Engagement and wellbeing. Much of the experimental evidence concerns social media, and it points both ways. People paid to deactivate Facebook for four weeks gained about an hour a day and reported slightly higher wellbeing, by 0.09 standard deviations, and afterwards used and valued the service less (Allcott et al., 2020). When people could set binding limits on their own screen time, their use of six apps fell by 16%, and the authors' model attributes 31% of social media use to self-control problems (Allcott, Gentzkow and Song, 2022). Against this, the median price at which participants in the deactivation study, weighted to resemble US Facebook users, would give Facebook up for four weeks was US$100 (Allcott et al., 2020), and an earlier choice experiment put it at about US$48 a month (Brynjolfsson, Collis and Eggers, 2019). Across three large surveys of adolescents, a correlational analysis found that technology use accounted for 0.4% or less of the variance in wellbeing (Orben and Przybylski, 2019). The strongest opposing view is the theory of rational addiction, in which even heavy, habit-forming use is what a forward-looking person plans (Becker and Murphy, 1988), so that time spent is evidence of value. Our claim, that time is the wrong target, does not need time spent to be harmful.

Formal models. Kleinberg, Mullainathan and Raghavan model a user with an impulsive and a reflective self, and show that a platform reading engagement as preference can lower the user's welfare even when welfare is its aim; they set incentives aside on purpose (Kleinberg, Mullainathan and Raghavan, 2022). Ichihashi and Kim study the incentives, with rational consumers and platforms that choose how addictive to be. A platform that earns from attention can gain by making its service more addictive, and a monopolist does so once attention is not scarce, while platforms that charge a price unrelated to attention choose zero addictiveness (Ichihashi and Kim, 2022). Our Proposition 2 is close to their pricing result, but there a price set against the consumer's surplus rewards zero addictiveness, and here a fixed price does so only through μ. Proposition 3 adds a cost of serving each minute, which separates a model on a server from one on a phone.

Companions. Evidence on companions themselves is recent, mostly in working papers and preprints, and does not agree. In fifteen-minute sessions, talking to an AI companion lowered momentary loneliness about as much as talking to a person, and did so again across a week-long study (De Freitas et al., 2024). In a four-week randomised trial of a general-purpose chatbot with 981 participants, the assigned voice and conversation conditions made little or no detectable difference to loneliness, socialising, emotional dependence or problematic use, but participants who chose to spend more time with the chatbot ended with worse scores on all four; time was not randomised (Fang et al., 2025). A survey of 1,131 users of one companion platform found use for companionship associated with lower wellbeing, again without a causal design (Zhang et al., 2025). Kirk and colleagues name the risk underneath: a system rewarded for longer conversations or better ratings may earn them at the cost of the person's longer-term wellbeing (Kirk et al., 2025).

03 Formulation

Take one person and one companion, and let t be the minutes the person spends with it in a day. Table 1 lists the notation. Looking back on the day, the person would say that those minutes were worth

$$U(t) = B(1 - e^{-t/\tau}) - \kappa t \tag{1}$$

The first term levels off. B is the most the companion can be worth to this person in a day, and τ is the time it takes to deliver about 63% of that. The second term charges κ for each minute, the value of whatever else the minute could have held. This is the form we used

for the length of a reply (Mitwa Lab, 2026a), with minutes in place of words. As there, it is an assumption, chosen as the simplest form with diminishing returns and a steady cost. When $B > \kappa\tau$, which we assume throughout, the person's best amount of time is

$$t^* = \tau \ln\left(\frac{B}{\kappa\tau}\right) \tag{2}$$

SYMBOL	MEANING
t, t^*	Minutes a day with the companion; the person's best amount
B, τ, κ	Most a day with it can be worth; time to deliver 63% of that; value of a minute spent elsewhere
a, $\bar{a}$, L	Pull of a design, per minute; its upper limit; value the person loses to it
r, c, n	Maker's revenue and serving cost per minute; $n = r - c$
μ, P	Rate at which the person's loss returns to the maker; income unrelated to time
F, h, σ	Subscription fee; effect on renewal of a minute of use, and of lost value
w_i, x_i	Weight and attainment of the person's i-th goal

Table 1. Notation.

Time actually spent need not equal t^*. In the moment, a person stays for another minute when staying feels worth more than leaving, and design changes how staying feels: a reply that ends on a question, a notification, a streak, an appeal at the moment of goodbye. We summarise all of this in one number, the *pull* a, in the same units as κ. A design with pull a adds a to the felt worth of each further minute, so the person stops when $(B/\tau)e^{-t/\tau} + a = \kappa$, which gives

$$t(a) = t^* + \tau \ln\left(\frac{\kappa}{\kappa - a}\right) \tag{3}$$

A positive pull holds the person past t^*. A negative pull is friction, such as slow replies or a daily cap, and it sends the person away early. The pull lies between $\kappa - B/\tau$, the friction at which nobody uses the product, and an upper limit $\bar{a}$, with $0 \le \bar{a} < \kappa$, the strongest pull that design can produce. This is a reduced form of the two-selves models in Section 2, with a as the extra value the impulsive self places on each further minute. With $a = 0$ the person stops where, on reflection, they would want to. That is a simplification. People have self-control problems of their own (Allcott, Gentzkow and Song, 2022), and in this note every departure from t^* is attributed to the maker. Call $U(t^*)$ the best value. A pull costs the person $L(a) = U(t^*) - U(t(a))$ of it, which works out to

$$L(a) = -\tau a - \kappa\tau \ln\left(1 - \frac{a}{\kappa}\right) \tag{4}$$

This is zero at $a = 0$ and positive otherwise. For a small pull it is close to $\tau a^2/2\kappa$, so time moves in proportion to a and value only in proportion to a^2.

04 Ways of being paid

The maker chooses the design, and with it the pull. Per person and per day, we write its profit as

$$\Pi(a) = P + n\,t(a) - \mu L(a), \qquad n = r - c \tag{5}$$

P collects income that does not depend on time, such as a flat fee or a one-time price. The maker earns r for each minute, for example from advertising shown during it, and pays c to serve it. The last term says that harming the person is not free. People who are worse off leave, complain and do not buy again, and $\mu \ge 0$ is the rate at which the person's loss comes back to the maker in those ways (a share, if value is measured in money). The linear form is an assumption, and Appendix A shows that the direction of each departure from t^* survives without it.

> **Proposition 1 (The pull a maker chooses).** *If $\mu > 0$, profit (5) is greatest at $a_M = n/\mu$, or at the nearer end of the feasible range when n/μ lies outside it. Inside the range the maker designs for the time in (6), which is the best amount of time for a person whose minutes cost $\kappa - n/\mu$. When $n > 0$ this exceeds t^* by $\tau \ln(\kappa/(\kappa - n/\mu))$. For $n > 0$ and fixed μ, that excess and the person's loss $L(n/\mu)$ both rise with n until n/μ reaches $\bar{a}$, after which the design stays at $\bar{a}$; the unconstrained rule would diverge as n/μ approaches κ.*

$$t_M = \tau \ln\left(\frac{B}{(\kappa - n/\mu)\tau}\right) \tag{6}$$

The proof is one derivative: $\Pi'(a) = \tau(n - \mu a)/(\kappa - a)$, which is positive below n/μ and negative above it. The maker sets the pull equal to what a minute earns it, divided by the rate at which the damage returns to it. A maker paid by the minute therefore treats the person's time as cheaper than it is, by n/μ, until the pull reaches $\bar{a}$, the most that design can do.

Figure 1 gives a sense of scale. With the illustrative values in its caption, a maker whose minute earns a quarter of $\mu\kappa$ designs for 22.9 minutes in place of 20, which is 14% more time at a cost to the person of 0.9% of the best value; at half of $\mu\kappa$ the figures are 26.9 minutes and 4.4%, and at nine tenths 43.0 minutes and 32%. In the same illustration a maker with $n = -0.5\mu\kappa$, the case of Proposition 3 below,

designs for 15.9 minutes, at a cost of 2.2%. None of these values has been fitted to anything.

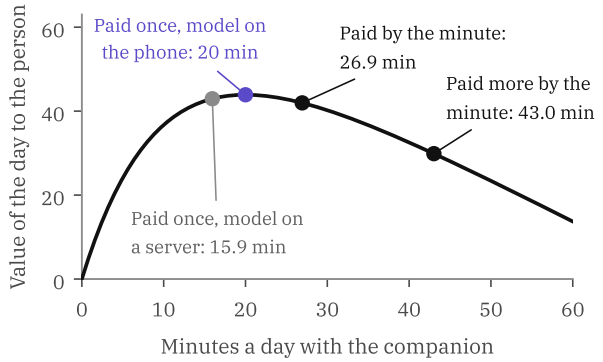


Figure 1. The value of a day with t minutes, from (1), and the time each maker designs for, from Proposition 1. Illustrative values: $\tau = 10$ minutes and $B = e^{2}\kappa\tau$, chosen to put the best amount at 20 minutes, with n/μ equal to -0.5κ, 0, 0.5κ and 0.9κ, and $\bar{a}$ above 0.9κ. Value is in units of the cost of one minute.

Proposition 2 (Paid once, model on the person's device).
If $r = 0$ and $c = 0$, no minute earns or costs the maker anything, and (5) becomes $\Pi(a) = P - \mu L(a)$. When $\mu > 0$ the maker's best design has no pull and the person spends t^{*}. When $\mu = 0$, profit equals P for every design, and the maker has no reason to prefer any amount of time to any other.

Proposition 3 (Paid once, model on a server).
If $r = 0$ and $c > 0$, the maker gains from less use. With $\mu > 0$, $a_M = -c/\mu$ and the maker designs for $t^{*} - \tau \ln(1 + c/\mu\kappa)$ minutes, or for none if that is negative, which is below t^{*}. With $\mu = 0$ it gains most if the product is not used at all. Minute for minute, for departures of equal length up to t^{*}, the person loses more from being sent away early than from being held late.

The sign has reversed. A paid-once product on the maker's servers is paid in advance and billed for every minute afterwards, so its interest lies in friction. The last sentence of the proposition is the asymmetry we proved for reply length (Mitwa Lab, 2026a), with the same caveat, set out in Section 9, that it depends on the form of (1). Putting the model on the person's device removes the cost, because replies then cost the maker nothing to produce (Mitwa Lab, 2026b).

Subscription. Where a subscription sits among these cases depends on what keeps people subscribed. Let the fee be F a period and the chance of renewal be

$$q(a) = q_0 + h\,[\,t(a) - t^{*}\,] - \sigma L(a) \tag{7}$$

Here $\sigma > 0$ says that people who lose value renew less, and $h \geq 0$ says that use in itself, as a habit, makes renewal more likely whether or not the time was well spent, with q_0 chosen so that $0 \leq q < 1$. If serving costs nothing, expected revenue $F/(1 - q)$ is greatest where q is, at

$$a_{\text{sub}} = \frac{h}{\sigma} \tag{8}$$

If renewal followed value alone, h would be zero and a subscription would give the same design as a one-time price. To the extent that renewal follows habit, the subscriber is pulled as an advertiser would pull them, with h/σ in the place of n/μ. Use of this kind appears to form habits, since in one experiment a temporary payment to cut social media use was still lowering it weeks after the payments stopped (Allcott, Gentzkow and Song, 2022). A cost of serving pushes the other way, as in Proposition 3. A club with a cost per visit earns most from the member who pays and stays away, and members of three US health clubs on a monthly fee who later cancelled went on paying, on average, for 2.31 full months after their last visit (DellaVigna and Malmendier, 2006). Renewal can also be raised without touching value or time, by making it hard to cancel (Mathur et al., 2019).

05 What a one-time price gets wrong

Proposition 2 concerns one incentive. A one-time price leaves other problems where they were and adds some of its own.

The first is timing. A companion is an experience good, one whose quality a buyer learns only by using it (Bergemann and Välimäki, 2005), and with a one-time price the whole payment is made before the experience. Suppose the number of buyers is $D(\hat{u})$, where $\hat{u} = (1 - \omega)s + \omega v$ is what a buyer expects the product to be worth: a mix of what the maker says, s, and the value that earlier buyers report, v. One more unit of reported value is then worth

$$m_{\text{once}} = P D'(\hat{u})\,\omega \tag{9}$$

to the maker, by analogy with μ, and one more unit of claimed value is worth $PD'(\hat{u})(1 - \omega)$. When $\omega < 1/2$ a unit of claimed value moves demand more than a unit of reported value, and when ω is small, because little reaches new buyers from old ones, it moves it much more. Read loosely, the reputational return that μ stands for in Proposition 2 is then small, and a one-time seller in that case has reason to press at the point of sale, as a maker paid by the minute has at the end of a conversation. A trial before purchase narrows the gap without closing it, since a few

minutes can show what a conversation is like but not what months are like.

The second is upkeep. After the sale, every update and fix is a cost with no revenue attached, and only μ rewards it. The third is that the maker may sell something else later. Mitwa has further models in preparation, so it has an interest in how Bloom's buyers regard it, and that interest is what makes μ positive. It is an incentive of its own as well, since regard and benefit can come apart. If buyers' later purchases rise with how much they used Bloom or how attached they became, at a given benefit, that channel enters profit as h does in (7) and, as in (8), restores a pull.

Last, the price is itself a cost. P does not affect the design that maximises (5), but US$200 paid before use puts the risk of a poor fit on the buyer.

So Proposition 2 should be read narrowly. A maker that is indifferent to a person's time lacks one bad incentive, and that absence is no evidence that it cares for the person or does them any good.

06 Why time cannot stand in for value

Even if minutes are not the maker's reward, they might still seem a fair measurement, since people spend more time on what helps them more. In the model this holds along one kind of variation and fails along another. Write $V = U(t)$ for the value a person gets from the time they spend.

> **Proposition 4 (Time does not identify value).** *(i) Across people who differ only in B, time and value rise together: $dV/dt = B/\tau - \kappa > 0$. (ii) Across designs that differ only in pull $a > 0$, time rises and value falls: $dV/dt = -a$. (iii) If both vary slightly and independently around a mean pull $a > 0$, the covariance of value and time, (10), can have either sign; around a mean pull $a \le 0$ it is non-negative to first order.*

$$\operatorname{Cov}(V, t) = \frac{\tau}{B}\left(1 - \frac{\kappa\tau}{B}\right)\operatorname{Var}(B) - \frac{\tau^2 a}{(\kappa - a)^2}\operatorname{Var}(a) \qquad (10)$$

Figure 2 draws the two paths. Part (i) is the opposing view of Section 2, and it is right as far as it goes: people with more to gain from a product use it more, and people who spend more time on Facebook also value it more (Brynjolfsson, Collis and Eggers, 2019). Part (ii) is the case this note is about, where the last minute bought by a pull a costs the person exactly a. A data set of people and their minutes mixes the two, so, where pull is present, the sign of the association depends on who is in it, which may be one reason the findings in Section 2 disagree. For the same reason, a maker that tests two designs which may differ both in what they deliver and in how hard they pull, and keeps the one with more minutes, cannot tell from minutes alone which is better for the person, unless it knows that neither design pulls.

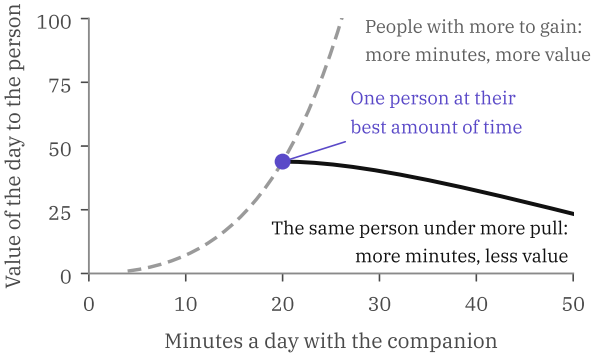


Figure 2. Two ways for minutes to rise, from Proposition 4, with the illustrative values of Figure 1. Along the dashed path people differ in how much they have to gain (B), with no pull. Along the solid path one person ($B = e^2\kappa\tau$) is held by pulls from 0 to 0.95κ.

07 Measuring care

What should be measured in place of minutes is whether the person got what they came for, by their own account. Two ideas from health care help, and we take them as measurement methods only. One is that the report comes from the person, with nobody else's reading laid over it (FDA, 2009). The other is that the person sets the terms. In goal attainment scaling a scale is set up for each of a person's goals, agreed with a clinician, before a service begins (Kiresuk and Sherman, 1968). Each goal has an expected level, scored 0, and outcomes run from −2 (much worse than expected) to +2 (much better). Goals can be weighted by how much they matter and how hard they are, or left unweighted (Turner-Stokes, n.d.). The scores combine into

$$T = 50 + \frac{10\sum_i w_i x_i}{\sqrt{(1-\rho)\sum_i w_i^2 + \rho\left(\sum_i w_i\right)^2}} \qquad (11)$$

where ρ is the assumed correlation between goals, by convention 0.3 (Turner-Stokes, n.d.). T is 50 when every goal is met as expected, and it is built to have a standard deviation of 10 (Turner-Stokes, 2009). To illustrate, with three equally weighted goals, exceeding one of them by a level gives 54.6.

For a companion, the analogue is an objective with no clock in it. Let a person name their own goals for the companion, in their own words: to think a decision through, to rehearse a hard conversation, to say what was on their mind before the day ends (these

examples are invented). A design is then scored by the weighted attainment it is expected to produce,

$$J = \mathrm{E}\Big[\sum_i w_i x_i\Big] \tag{12}$$

If attainment rises and falls with the value in (1), a maker rewarded on J alone is in the first case of Proposition 2 ($n = 0$, $\mu > 0$). Time enters J only through what it does for the goals. Some cautions belong with (11) and (12). Whoever sets the expectations can inflate T by setting them low, and practitioners are advised to check that the mean across people stays near 50 (Turner-Stokes, n.d.). The validity of the scale is unsettled (Gaasterland et al., 2016). A person's stated goals can also be shaped by the very system being scored against them (Kirk et al., 2025).

We cannot compute J for Bloom in use. Whatever goals a person states to Bloom, and what came of them, sit in conversations and memories, which are never sent to the lab. What is open to us is study by consent, with people who agree in advance to name goals and report on them. As an illustration, if T has its nominal standard deviation of 10, the usual two-sample formula asks for 63 people a group to detect a difference of 5 points between two designs, and 393 to detect 2, at 80% power and a two-sided 5% level.

08 Result

This note has no measurement to report. Bloom has been live on the iPhone since 11 June 2026, and Table 2 places it in the model using the lab's stated facts.

STATED FACT	IN THE MODEL	WHAT FOLLOWS
Paid once, US$200	Income is P alone	No renewal; but see later sales (Section 5)
No advertising	$r = 0$	A minute earns nothing
Runs on the iPhone	$c = 0$	A minute costs nothing to serve
Conversations and memories are never sent to Mitwa	J unobserved	Care cannot be scored from what is said

Table 2. Bloom in the model. The left column holds facts the company states about the product; Section 9 sets the second row beside the App Store label. The other columns are derived from them and were not measured.

With $r = 0$ and $c = 0$, Proposition 2 applies. Within (5), Mitwa's income from a person who has bought Bloom does not rise or fall with the minutes they spend, so the incentive (5) gives it to inflate those minutes is zero, unless people's later purchases of Mitwa's other models rise with how much they used Bloom or how attached they became rather than with what it did for them, which would act like h in (7) (Section 5). The zero is derived from Table 2 and (5), conditional on the stated facts and on that channel being absent, and it says nothing about behaviour.

A remaining cost per minute would make c slightly positive, which by Proposition 3 moves the incentive towards less use.

What has not been shown is longer. We have not shown that Bloom's design has no pull. Pull can arise in a conversational system without anyone choosing it, for instance where a system is rewarded for approval (Kirk et al., 2025) or where behaviour is read as preference (Kleinberg, Mullainathan and Raghavan, 2022), and a maker with no reason to add pull also has no revenue reason to look for it. We do not know μ for Mitwa, or t^* for anyone, and this note uses no figure for how long people spend with Bloom. Whether people return to it over a year is not yet known, and by Section 6 returning would not settle whether it serves them. We have no evidence that time with Bloom is good for those who spend it.

09 Limitations and broader impact

There are no data in this note. Every number in Sections 4 to 7 that is not quoted from a cited study is an illustration, and none describes Bloom or the people who use it.

The model is stylised. It has one person, one companion and one number for design. Real designs differ in kind as well as strength, and a pull that is harmless for one person may not be for another.

The functional forms (1) and (5) are assumptions, and nothing has been fitted to data. The direction of the departure from t^* in Propositions 1 to 3 needs only that the person's value has a single best point and that the maker's profit does not fall when the person is better off (Appendix A). The sizes, the rule $a_{\mathrm{M}} = n/\mu$, the asymmetry in Proposition 3 and the growth of the loss with n depend on the forms; as Note 02 says of reply length, a cost per minute that rises with time could weaken or reverse the asymmetry.

Real products mix ways of being paid, and the apps in the audit of Section 1 do not sort neatly by revenue; the one with the lowest rate among the five that sent such messages carries advertising (De Freitas, Oğuz-Uğuralp and Kaan-Uğuralp, 2025). Proposition 1 has not been tested, here or by that audit.

Reputation is outside the model. μ stands for everything that makes a maker answerable to people it has already sold to, and we treated it as a constant.

Our statements about what the lab can see are limited to what is said in conversations and what Bloom remembers, which are never sent to Mitwa; voice-to-text, if a person uses it, can go to Apple's speech service. The App Store listing for the Mitwa AI app declares that usage data and advertising data may be collected and linked to the person's identity, for the developer's advertising or marketing, for analytics and for product personalisation, and that diagnostic data may be collected without being linked (Apple, 2026). This note cannot reconcile that label with the lab's stated facts that Bloom is paid for once and carries no advertising. The propositions do not depend on the label, but the result for Bloom does. Placing Bloom at $r = 0$ in Table 2, and so the zero incentive of Section 8, rests on the stated facts, and would fail if any of Mitwa's revenue rose with use.

Nothing here shows that Bloom is good for anyone. A zero incentive to waste a person's time is compatible with a product that wastes it by accident, and with one that is simply of little use.

On broader impact, we think the way a companion is paid for deserves the same scrutiny as the way it is built, because it decides what its maker is rewarded for noticing. We would rather be paid in a way that makes a person's minutes worthless to us. That choice has the costs listed in Section 5, and it leaves the case for benefit to measurement we have not yet done. This note makes no clinical claim.

10 What a proper evaluation would report

An evaluation of whether a companion serves people should report at least these.

- How the maker is paid, and whether any of its revenue or cost moves with time spent (n in (5)).
- Time spent as a description, never as the outcome, beside the time each person says they would choose on reflection.
- The share of goodbyes that the product answers with an appeal to stay, audited as in De Freitas, Oğuz-Uğuralp and Kaan-Uğuralp (2025).
- Goals and expected levels set by each participant before use, and attainment by (11), with the mean of T as a check on the expectations.
- A comparison in which time is assigned, since chosen time is confounded with how much a person has to gain (Proposition 4).
- Outcomes at a session, a month and a year, since findings over single sessions and over four weeks have differed, and no study cited here runs to a year.
- How consent was obtained, and by what route any of this left the device.

11 Conclusion

A companion can be paid for in ways that make a person's minutes valuable to its maker, or in a way that makes them worthless. In our model, the first kind of maker designs for more time than the person would choose, and the excess grows with what a minute earns, up to the limit of what design can do. A maker paid once, with the model on the person's phone, earns nothing from a minute and has no reason to add one, unless its later sales reward attachment more than benefit. By the lab's stated facts, Bloom is sold the second way.

That is all the arrangement buys. It does not show that Bloom has no pull, that its makers attend to people who have already paid, or that an hour with it is an hour well spent. Showing those things needs outcomes that people report against goals they set, gathered with their consent, because the conversations themselves never reach us.

Disclosure. Bloom is a product of Mitwa AI, and this note was written by the lab that builds it. The note finds an advantage in the way Bloom is sold, a conclusion in which its authors have a commercial interest; Section 5 should be weighed with that in mind. No user conversation is available to us, and none was used. Reproducibility. Every result in Sections 3 to 7 is closed-form and proved in the text or in Appendix A, and each proposition was checked numerically over a grid of values by a script kept with the note.

References

Allcott, H., Braghieri, L., Eichmeyer, S., and Gentzkow, M. (2020). The welfare effects of social media. American Economic Review, 110(3), 629–676. doi.org/10.1257/aer.20190658

Allcott, H., Gentzkow, M., and Song, L. (2022). Digital addiction. American Economic Review, 112(7), 2424–2463. doi.org/10.1257/aer.20210867

Apple (2026). Mitwa AI. App Store listing, app privacy section, India storefront. Opened 12 November 2026. apps.apple.com/in/app/mitwa-ai/id6755034468

Becker, G. S., and Murphy, K. M. (1988). A theory of rational addiction. Journal of Political Economy, 96(4), 675–700. doi.org/10.1086/261558

Bergemann, D., and Välimäki, J. (2005). Monopoly pricing of experience goods. Cowles Foundation Discussion Paper 1463R, Yale University. cowles.yale.edu/sites/default/files/2022-08/d1463-r.pdf

Brynjolfsson, E., Collis, A., and Eggers, F. (2019). Using massive online choice experiments to measure changes in well-being. Proceedings of the National Academy of Sciences, 116(15), 7250–7255. doi.org/10.1073/pnas.1815663116

De Freitas, J., Oğuz-Uğuralp, Z., and Kaan-Uğuralp, A. (2025). Emotional manipulation by AI companions. Harvard Business School Working Paper 26-005. arxiv.org/abs/2508.19258

De Freitas, J., Uguralp, A. K., Uguralp, Z. O., and Puntoni, S. (2024). AI companions reduce loneliness. Harvard Business School Working Paper 24-078. arxiv.org/abs/2407.19096

DellaVigna, S., and Malmendier, U. (2006). Paying not to go to the gym. American Economic Review, 96(3), 694–719. doi.org/10.1257/aer.96.3.694

Fang, C. M., Liu, A. R., Danry, V., Lee, E., Chan, S. W. T., Pataranutaporn, P., Maes, P., et al. (2025). How AI and human behaviors shape psychosocial effects of extended chatbot use: A longitudinal randomized controlled study. arxiv.org/abs/2503.17473

FDA (2009). Guidance for industry. Patient-reported outcome measures: Use in medical product development to support labeling claims. U.S. Food and Drug Administration. www.fda.gov/media/77832/download

Gaasterland, C. M. W., Jansen-van der Weide, M. C., Weinreich, S. S., and van der Lee, J. H. (2016). A systematic review to investigate the measurement properties of goal attainment scaling, towards use in drug trials. BMC Medical Research Methodology, 16, 99. doi.org/10.1186/s12874-016-0205-4

Ichihashi, S., and Kim, B.-C. (2022). Addictive platforms. Bank of Canada Staff Working Paper 2022-16. doi.org/10.34989/swp-2022-16

Kiresuk, T. J., and Sherman, R. E. (1968). Goal attainment scaling: A general method for evaluating comprehensive community mental health programs. Community Mental Health Journal, 4(6), 443–453. doi.org/10.1007/BF01530764

Kirk, H. R., Gabriel, I., Summerfield, C., Vidgen, B., and Hale, S. A. (2025). Why human-AI relationships need socioaffective alignment. arxiv.org/abs/2502.02528

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2022). The challenge of understanding what users want: Inconsistent preferences and engagement optimization. arxiv.org/abs/2202.11776

Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., and Narayanan, A. (2019). Dark patterns at scale: Findings from a crawl of 11K shopping websites. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 81. doi.org/10.1145/3359183

Mitwa Lab (2026a). Eleven words: Reply length in AI companions. Technical note 02.

Mitwa Lab (2026b). A companion in 2 GB. Technical note 03.

Orben, A., and Przybylski, A. K. (2019). The association between adolescent well-being and digital technology use. Nature Human Behaviour, 3, 173–182. doi.org/10.1038/s41562-018-0506-1

Turner-Stokes, L. (2009). Goal attainment scaling (GAS) in rehabilitation: A practical guide. Clinical Rehabilitation, 23(4), 362–370. doi.org/10.1177/0269215508101742

Turner-Stokes, L. (n.d.). Goal attainment scaling (GAS) in rehabilitation: A practical guide. King's College London. www.kcl.ac.uk/nmpc/assets/rehab/6.-goal-attainment-scaling-a-practical-guide.pdf

Zhang, Y., Zhao, D., Hancock, J. T., Kraut, R., and Yang, D. (2025). Interaction with AI companions and psychological well-being. arxiv.org/abs/2506.12605

A Appendix A: Proofs

Equations (2) to (4).

$U'(t) = (B/\tau)e^{-t/\tau} - \kappa$ is decreasing, so (1) is concave with its maximum where $Be^{-t/\tau} = \kappa\tau$, which is (2). The stopping rule $(B/\tau)e^{-t/\tau} = \kappa - a$ gives $t(a) = \tau \ln(B/((\kappa - a)\tau))$, which is (3). Then $U(t(a)) = B - (\kappa - a)\tau - \kappa t(a)$ and $U(t^*) = B - \kappa\tau - \kappa t^*$. Subtracting, and using (3) for $t(a) - t^*$, gives (4). With $x = a/\kappa$, $L = \kappa\tau f(x)$ with $f(x) = -x - \ln(1 - x)$. As $f(0) = 0$ and $f'(x) = x/(1 - x)$, $L > 0$ for every feasible $a \neq 0$. Near zero $f(x) \approx x^2/2$.

Proposition 1.

From (3) and (4), $t'(a) = \tau/(\kappa - a)$ and $L'(a) = \tau a/(\kappa - a)$, so $\Pi'(a) = nt'(a) - \mu L'(a) = \tau(n - \mu a)/(\kappa - a)$. For $a < \kappa$ its sign is that of $n - \mu a$, so Π rises up to $a = n/\mu$ and falls after it, and on a restricted range the maximum is at the nearer end. Substituting $a = n/\mu$ in (3) gives (6). The excess $\tau \ln(1/(1 - x))$ and the loss $\kappa\tau f(x)$ at $x = n/\mu\kappa$ increase with n for $n > 0$ while $n/\mu < \bar{a}$; beyond that the pull stays at $\bar{a}$, and only the unclipped forms diverge as x tends to 1. For $n < 0$ the loss falls as n rises, so the loss rises with n only for $n > 0$.

Propositions 2 and 3.

Set $n = 0$ in Proposition 1 for the first (zero pull is feasible as $0 \leq \bar{a}$), and $n = -c$ with $\mu > 0$ for the second. If $\mu = 0$ and $n = 0$, (5) is the constant P. If $\mu = 0$ and $n = -c$, (5) is $P - ct$, greatest at $t = 0$. For the asymmetry, with $0 < d \leq t^*$, $(B/\tau)e^{-t/\tau} = \kappa$ at t^*, so a visit d minutes short of t^* loses $\kappa[\tau(e^{d/\tau} - 1) - d]$ and one d minutes over loses $\kappa[d - \tau(1 - e^{-d/\tau})]$. The difference is $2\kappa[\tau \sinh(d/\tau) - d] > 0$. This is per minute; per unit of pull, $L(a) > L(-a)$.

Equations (8) and (9).

$q'(a) = ht'(a) - \sigma L'(a) = \tau(h - \sigma a)/(\kappa - a)$, which changes sign from positive to negative at h/σ. With a serving cost the maker maximises $(F - kct)/(1 - q)$, k the days in a period; when $F - kct > 0$ its derivative is negative at and above h/σ, so the chosen pull is lower. Later income that rises with time at a given value works as h does. For (9), revenue is $PD(\hat{u})$, with derivatives $PD'(\hat{u})\omega$ in v and $PD'(\hat{u})(1 - \omega)$ in s.

Proposition 4.

With pull a, $V = B - (\kappa - a)\tau - \kappa t$. Varying B: $\mathrm{d}t/\mathrm{d}B = \tau/B$ and $\mathrm{d}V/\mathrm{d}B = 1 - \kappa\tau/B$, whose ratio is $B/\tau - \kappa$. Varying a: $\mathrm{d}t/\mathrm{d}a = \tau/(\kappa - a)$ and $\mathrm{d}V/\mathrm{d}a = -\tau a/(\kappa - a)$, whose ratio is $-a$. For small independent changes around the mean point, the first-order covariance is each product of derivatives times its variance, which is (10). Its first term is positive as $B > \kappa\tau$, and its second has the sign of $-a$, so (10) can be negative only when $a > 0$. At $a > 0$ it is positive when only B varies and negative when only a varies, which gives part (iii).

Directions without the functional forms.

Suppose only that U has a unique maximum at t^*, where it is differentiable, and that profit is $P + nt + G(U(t))$ with G differentiable and nondecreasing. If $n > 0$, every $t < t^*$ gives less profit than t^*, and the derivative of profit at t^* is $n > 0$, so any profit-maximising time is above t^*. If $n < 0$ the same argument puts it below. If $n = 0$, t^* maximises profit, uniquely when G is increasing, and every t does when G is constant. This argument concerns only the sign of $t - t^*$. It does not give the sizes, the asymmetry in Proposition 3 or the growth of the loss with n, and it assumes that a maximiser exists.