Chopra, Felix; Haaland, Ingar K.; Roever, Nicolas; Roth, Christopher

Working Paper

Evaluating Behavioral Interventions at Scale with AI

CESifo Working Paper, No. 12410

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Chopra, Felix; Haaland, Ingar K.; Roever, Nicolas; Roth, Christopher (2026) : Evaluating Behavioral Interventions at Scale with AI, CESifo Working Paper, No. 12410, Munich Society for the Promotion of Economic Research - CESifo GmbH, Munich

This Version is available at:
https://hdl.handle.net/10419/338374

CES ifo
Working Papers

12410
2026

January 2026

Evaluating Behavioral Interventions at Scale with AI

Felix Chopra, Ingar Haaland, Nicolas Roever, Christopher Roth

CES ifo

Imprint:

CESifo Working Papers

ISSN 2364-1428 (digital)

Publisher and distributor: Munich Society for the Promotion
of Economic Research - CESifo GmbH

Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740

Email office@cesifo.de
https://www.cesifo.org

Editor: Clemens Fuest

An electronic version of the paper may be downloaded free of charge
· from the CESifo website: www.ifo.de/en/cesifo/publications/cesifo-working-papers
· from the SSRN website: www.ssrn.com/index.cfm/en/cesifo/
· from the RePEc website: https://ideas.repec.org/s/ces/ceswps.html

Evaluating Behavioral Interventions at Scale with AI

Felix Chopra Ingar Haaland Nicolas Roever Christopher Roth

January 17, 2026

Abstract

We test the effectiveness of different AI-delivered conversation protocols to increase people's motivation for change. In a large-scale experiment with 2,719 social media users, we randomly assign participants to a control conversation or one of three treatment arms: two Motivational Interviewing protocols promoting self-persuasion (change focus or decisional balance) and a direct persuasion protocol providing unsolicited advice and information. All conversations are led by an AI interviewer, enabling standardized delivery of each protocol at scale. Our results show that all three interventions significantly increase motivation for change and the perceived costs of social media use, with change-focused self-persuasion yielding the largest effects. These effects persist and translate into self-reported reductions in social media use more than two weeks after the intervention. Our findings illustrate how AI-led conversations can serve as a scalable platform both for delivering behavioral interventions and for testing what makes them effective by systematically varying how conversations are conducted.

Keywords: AI interviews, Scaling, Motivation, Persuasion, Social Media, Beliefs.
JEL codes: C90, D83, D91

Chopra: Frankfurt School of Finance & Management, CESifo, f.chopra@fs.de; Haaland: NHH Norwegian School of Economics, FAIR, CEPR, NTNU, Ingar.Haaland@nhh.no; Roth: University of Cologne and ECONtribute, Max Planck Institute for Behavioral Economics, CEPR, NHH, roth@wiso.uni-koeln.de; Roever: University of Cologne, nicolas.roever@wiso.uni-koeln.de. We would like to thank Andreas Grunewald, Fabian Roeben, and Frederik Schwerter for valuable feedback. Moritz Lakenbrink, Maximilian Müller, and Rosanna Simonis provided excellent research assistance. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/2-390838866. Haaland acknowledges financial support from the Research Council of Norway (RCN) through the project "Media Bias and Political Polarization," under Grant Number 344979. Haaland and Roth acknowledge financial support from the RCN through its Centre of Excellence Scheme (FAIR project No 262675). The data collections were pre-registered at AsPredicted (#264681 and #266382). Ethics approval was obtained from the IRB boards at the Frankfurt School of Finance & Management, NHH Norwegian School and the University of Cologne.

1 Introduction

A central question in the social sciences is how to effectively increase people's motivation to change their behavior in domains where they experience self-control problems. From excessive social media use to lack of exercise and undersaving for retirement (Allcott et al., 2022; Benartzi and Thaler, 2007; DellaVigna and Malmendier, 2006), many of these behaviors persist not because people do not know what is good for them or lack the intention to change—but rather because they fail to act on their good intentions (Godin and Conner, 2008). Understanding which types of interventions can most successfully motivate change—and through which mechanisms they operate—remains a core open question.

Two starkly different approaches to induce changes in motivation are ones that provide clear advice and information ("direct persuasion") and approaches in which individuals themselves generate arguments for change ("self-persuasion"). While economists typically focus on direct persuasion (DellaVigna and Gentzkow, 2010), reactance theory from psychology predicts that attempts at direct persuasion will often backfire by creating psychological resistance (Brehm, 1966). Clinical psychologists therefore typically focus on self-persuasion techniques to increase people's motivation for change, such as Motivational Interviewing (short: MI; Miller and Rollnick, 2012). In these interventions, the focus is on eliciting people's own reasons for change rather than telling them what to do.

Although MI is widely used in clinical and applied settings, evidence on its effectiveness is mixed (Burke et al., 2003; Frost et al., 2018; Zhu et al., 2024). Randomized trials of MI are often small-scale and bundled with other interventions (Frost et al., 2018; Zhu et al., 2024). There is also very limited experimental evidence on the effectiveness of different conversation protocols, limiting our knowledge of the potential mechanisms contributing to behavioral change (Apodaca and Longabaugh, 2009). For instance, there is an ongoing debate about whether self-persuasion interventions should focus on evoking pro-change change arguments or systematically explore ambivalence by treating pros and cons in a balanced way (Miller and Rose, 2015). There is also mixed evidence on the success of direct persuasion, with conflicting evidence from economics and psychology (DellaVigna and Gentzkow, 2010; Steindl et al., 2015).

Progress on these questions has been limited by several limitations of the existing evidence. First, interventions randomizing conversation protocols are often infeasible to implement at scale. Second, conversation protocols are difficult to standardize with human interviewers. Third, many studies fail to document the integrity of the MI delivery, making it unclear whether they even qualify as MI interventions according to established standards (Frost et al., 2018). This paper addresses these limitations by leveraging an LLM-based approach that allows us to experimentally compare the efficacy of distinct conversation protocols at scale, documenting adherence to MI protocols, and holding constant tone, length, and delivery. We focus on social media consumption, recognizing that many people are addicted to social media and want to

reduce their consumption (Allcott et al., 2022) or would even prefer to live in a world without it (Bursztyn et al., 2025), consistent with studies showing negative mental health effects of social media use (Allcott et al., 2020; Braghieri et al., 2022; Haidt, 2024).

Design In a large-scale experiment with 2,719 social media users recruited online, we systematically vary whether conversations rely on self-persuasion or direct persuasion, and whether the treatments aimed at self-persuasion focus on change or exploring ambivalence. In the first treatment (*Change Talk*), the AI interviewer is instructed to reinforce change talk and redirect away from arguments favoring the status quo. In the second treatment (*Decisional Balance*), the AI interviewer follows a similar MI protocol except that they are instructed to explicitly explore ambivalence by treating advantages and disadvantages of change in a balanced way instead of trying to steer the conversation towards change talk. In the third treatment (*Direct Persuasion*), we deviate from all standard MI protocols and design an AI interviewer that engages in direct persuasion to make people more motivated to reduce their social media use. This treatment combines information provision with a clear stance that respondents *should* reduce their consumption and also provides respondents with unsolicited advice and guidance on how to achieve this target. Finally, respondents assigned to an active control group face a neutral interview about their time use. This design enables a clean comparison of leading theories of motivational change within a unified experimental environment.

While the treatments differ in their focus—change-focused self-persuasion, self-persuasion with decisional balance, and direct persuasion—the treatments were designed to be similar in length and have a roughly identical structure. Across all treatments, the interview starts with an exploration of the respondents' current social media habits and then moves on to "scaling questions," in which respondents are asked how motivated they are to reduce their social media time on a scale from 0 to 10. After the scaling question and the associated discussion of commitment and confidence in the ability to implement change, the interview ends with a planning section on how to achieve the goals.

When the interviews are finished, we inform respondents that they will be asked some questions about their social media consumption. In particular, we measure motivation for change, a battery of questions on costs, benefits, self-efficacy, awareness, ideal use, and intended time use, and an incentive-compatible elicitation of willingness to pay for a third-party app blocker. Around two weeks after the main experiment, we conduct a follow-up study to examine whether the intervention had persistent effects on motivation and beliefs and whether it affected self-reported social media time.

Validation of AI interviews A central challenge in evaluating conversational interventions, and MI in particular, is documenting (i) high fidelity of treatment implementation and (ii) systematic differences in interviewer behavior at scale. We address this challenge by combining participant evaluations with large-scale transcript analysis, leveraging LLMs to measure fidelity

in a standardized and scalable way.

We validate the implementation in three ways. First, participants' evaluations of interviewer behavior, measured using the Client Evaluation of Counseling Scale (Madson et al., 2013), differ across treatment arms in line with the experimental design. *Change Talk* interviews are perceived as more collaborative and autonomy-supportive, *Decisional Balance* interviews as more balanced and exploratory, and *Direct Persuasion* interviews as more directive and authoritative.

Second, to assess adherence to motivational interviewing principles, we develop an LLM-based pipeline to score interview transcripts using the Motivational Interviewing Treatment Integrity (MITI) coding manual—the gold standard for documenting MI fidelity (Moyers et al., 2016). We validate our LLM-based measurement on a ground truth of expert-annotated human-led motivational interviews. Applying this pipeline to our study transcripts, we find that both MI treatments score highly on MITI measures of MI quality. Moreover, the distribution of quality scores is strongly concentrated, documenting the high consistency in AI delivery.

Third, we examine whether conversational content differs systematically across protocols using an embedding-based topic model. *Decisional Balance* interviews more frequently explore trade-offs between costs and benefits, while *Direct Persuasion* places greater emphasis on concrete implementation strategies. Taken together, these results indicate that the AI interviews were implemented as intended and generated distinct conversational experiences, providing a meaningful first-stage for analyzing their causal effects on motivation, beliefs, and behavior.

Main results Our main result is that *Change Talk* and *Direct Persuasion* are both highly effective at increasing motivation for change, leading to increases in motivation to reduce smartphone use by 0.52 and 0.43 standard deviations, respectively, compared to the control group assigned to neutral time use interviews (both $p < 0.001$). Conversations seeking decisional balance also increase motivation significantly but are less effective than the other two treatments (0.21 standard deviations, $p < 0.01$). These effects persist: in our two-week follow-up survey, participants in *Change Talk* and *Direct Persuasion* still report 0.15 and 0.16 standard deviations higher motivation (both $p < 0.01$), while for *Decisional Balance* the effects are somewhat more muted (0.07 standard deviations) and not statistically significant at conventional levels.

While motivation captures a key psychological precondition for behavior change, it may also depend on the perceived costs of that effort. Here, we find striking differences across treatments. *Change Talk* is by far the most effective in shifting perceived costs and benefits of social media use, increasing perceived net costs by 0.45 standard deviations, while *Decisional Balance* and *Direct Persuasion* increase perceived costs by 0.13 and 0.20 standard deviations, respectively. These differences persist in our follow-up survey for *Change Talk* and *Direct Persuasion*, while the effect of *Decisional Balance* fades to statistical insignificance.

A natural concern with self-reported motivation and belief measures is that they could reflect social desirability bias or experimenter demand effects. To address this, we examine willingness

to pay for a premium version of an app blocker using an incentivized measure, where choices carry actual financial consequences. We find a large and significant treatment effect for *Change Talk*, increasing willingness to pay by \$0.82 (27% relative to the control mean of \$3.00, $p < 0.01$). *Decisional Balance* and *Direct Persuasion* also generate positive treatment effects on willingness to pay, but the point estimates are smaller and not statistically significant.

Does the increased motivation for change among treated respondents also translate into reductions in actual social media use? We first examine ideal and predicted time use. Participants in *Change Talk* and *Direct Persuasion* report wanting to spend 7–8 fewer minutes per day on social media ($p < 0.01$ and $p < 0.05$, respectively). More strikingly, participants predict substantial reductions in their future use: 25 minutes per day for *Change Talk*, 14.5 minutes for *Decisional Balance*, and 39 minutes for *Direct Persuasion*—the largest predicted reduction (all $p < 0.001$). Turning to self-reported social media use in our two-week follow-up, we find that all treatments reduce time spent on social media relative to control. Notably, *Direct Persuasion* generates the largest behavioral effects at 24 minutes per day less ($p < 0.001$), compared to 11–12 minutes for *Change Talk* and *Decisional Balance* (both $p < 0.10$). Furthermore, a strong and robust correlation between self-reported and verified social media time for a subsample of respondents who uploaded screenshots alleviates potential concerns about reporting biases.

How do people manage to reduce their social media use? To answer this, we elicit the strategies that participants relied upon in our follow-up survey. We first document that participants are more likely to use strategies discussed during their conversation with the AI interviewer. We then group strategies into two categories: behavioral strategies (e.g., setting specific rules and goals, relying on willpower) and technology-based strategies (e.g., using app blocker, changing phone settings). While participants across all three treatments report higher use of behavioral strategies to reduce their social media time, we observe the highest adoption rate of technology-based strategies among respondents in the *Direct Persuasion* treatment. This pattern can help explain why the *Direct Persuasion* treatment—where the interviewer puts emphasis on creating concrete strategies for translating intentions into action—induces the largest effects on social media use use despite being marginally less effective than *Change Talk* at raising people's motivation to change.

Literature review Our paper relates to several strands of the literature. First, we contribute to the literature on behavioral interventions to motivate change. Within this literature, there is an ongoing discussion on how interventions aimed at *self-persuasion*, such as Motivational Interviewing (MI) techniques (Miller and Rollnick, 2013), should treat ambivalence towards change. One common technique, Decisional Balance (DB), involves exploring ambivalence by going through the advantages and disadvantages of behavioral change in a balanced way. The basic idea is that a DB approach helps reduce decision errors and makes plans in favor of behavioral change more robust to subsequent conflict and stress (Janis and Mann, 1977). While DB was considered a key aspect in early formulations of MI (Miller and Rollnick, 1991), current

textbook formulations instead favor a change-focused approach. In this approach, the interviewer seeks to evoke pro-change arguments ("change talk") while responding to, but not amplifying, arguments against change ("sustain talk") (Miller and Rose, 2015). One of the arguments for avoiding DB in Motivational Interviews is evidence from studies showing that a higher ratio of change talk to sustain talk is a good predictor of change (Moyers et al., 2009), though evidence from recent meta-studies provides conflicting evidence on whether a systematic evaluation of pros and cons predicts positive or negative outcomes (Mair et al., 2023; Samdal et al., 2017).

While the most credible way to understand the effectiveness of MI and the mechanisms behind successful behavioral change is through experimental evidence, randomized trials of MI are often small-scale and bundled with other interventions (Frost et al., 2018; Zhu et al., 2024), leading to a lack of credible evidence on mechanisms and overall effectiveness. Results from experimental studies are also mixed. LaBrie et al. (2006) study the effect of a DB procedure within an MI interview and conclude that DB plays an important role in MI and could be an effective intervention in itself. By contrast, Carey et al. (2006) show that a brief MI reduced drinking problems among college students, whereas an enhanced MI that incorporated a decisional balance component did not. Their evidence suggests that the decisional balance intervention inadvertently reactivated the perceived benefits of drinking, thereby attenuating the change motivation generated by the MI. However, a subsequent study by Foster et al. (2015) finds that a DB intervention was effective in reducing college drinking, underscoring the mixed nature of these experiments. Our results show that MI can be effective irrespective of whether it focuses on change or resolving ambivalence through decisional balance, although change-focused MI seems to be more effective overall by not focusing attention on the perceived positive aspect of the status quo.

While self-persuasion protocols differ on how to deal with ambivalence, the core unifying point is to avoid direct and unsolicited attempts at persuasion and instead allow the interviewee to make up their own reasons for change. These insights build on reactance theory from psychology in which attempts at persuasion—especially attempts that use forceful language—can trigger perceived threats of freedom and lead to *lower* motivation for change (Brehm, 1966; Steindl et al., 2015). By contrast, our findings suggest that direct attempts at persuasion—which violate MI protocols by providing unsolicited information, taking a clear stance that change is needed, and providing possible solutions—can be equally effective as a change-oriented MI. This evidence aligns with broader evidence from economics, which shows that direct attempts at persuasion often succeed in changing beliefs and attitudes (DellaVigna and Gentzkow, 2010; Haaland et al., 2023).

Moreover, our paper contributes to a growing literature in economics using large language models to collect and analyze qualitative data at scale (Ash and Hansen, 2023; Braghieri et al., 2025; Chopra and Haaland, 2023; Galasso et al., 2024; Geiecke and Jaravel, 2025; Graeber et al., 2024; Haaland et al., 2025; Habibi et al., 2024). We contribute to this body of work by

documenting how the conversational records from AI-delivered interventions can be used to examine the quality and fidelity of treatment implementation.

Finally, our work relates to an emerging literature on the scaling of behavioral interventions (Al-Ubaydli et al., 2017; DellaVigna and Linos, 2022; List, 2022, 2024) and an ongoing debate on the role of algorithms and AI in shaping the scaling of behavioral interventions (Mullainathan and Rambachan, 2025) and scientific discovery more broadly (Ludwig et al., 2025; Mullainathan, 2025). Particularly related to our study are attempts to look at the effects of LLM-based chatbots (Ash et al., 2025) or apps (Chopra et al., 2025) in the context of political beliefs. For example, Costello et al. (2024) show that LLM-based chatbots can durably reduce conspiracy beliefs, even if the AI is perceived as a human (Boissin et al., 2025). Our contribution to this literature is twofold. First, we provide a proof of concept that AI can be effectively used to scale up the delivery of conversation-based treatment protocols that are commonly used by practitioners and field workers, such as motivational interviewing or cognitive behavioral therapy (Blattman et al., 2017). Our second contribution is to provide a template for testing hypotheses about what precisely makes conversation-based treatment protocols effective by varying their individual components in a large-scale trial in a systematic way.

2 Experimental design and data

In this section, we provide details on the sample and the experimental design. Appendix Section E provides the full experimental instructions and Figure B.1 presents an overview of the design.

2.1 Sample

We recruited 2,800 participants on Prolific, a widely used platform for social science research, between December 18 and 20, 2025. We restricted recruitment to UK and US Prolific users who had indicated having Instagram or TikTok installed on their smartphone prior to the survey.

To mitigate concerns about bots and LLM use by participants (Rilla et al., 2025), we exclude participants that type unusually fast (more than 10 characters per second) and those that fail an audio screener that requires participants to follow instructions recorded in an audio file. Celebi et al. (2025) find that these types of screeners effectively detect over 90% of AI bots, and that AI bots are mostly absent on Prolific. We additionally exclude the fastest and slowest 1% of participants based on pre-treatment survey time to increase data quality. After applying these preregistered filters, our final analysis sample consists of 2,719 participants.

Column 1 of Table A.1 provides summary statistics for our final sample and columns 2–7 show that our sample is balanced on observables across treatment arms, thus suggesting that the randomization worked as expected. 47% of our participants reside in the US and 53% in the UK. In terms of age, our sample closely tracks the average age of the general population (US: 42; UK: 41). Similarly, participants' household income of $80,270 is close to the average

income in the US ($81,000) and the UK ($74,300). As is common in online surveys (Armantier et al., 2017), our sample has a higher share of female (60%) and college-educated participants (62%) compared to the general population. Consistent with our recruitment strategy, 92% of our final sample reports being an Instagram or TikTok user, which is substantially above the general population shares of 50% for Instagram and 37% for TikTok in the US (Auxier and Anderson, 2021). Notably, participants spend approximately three hours (183 minutes) on average per day on social media apps (standard deviation: 111 minutes). 85% of the sample spends at least an hour per day on social media.[1] This underscores that we were able to recruit the relevant population of heavy social media users.

2.2 Survey

2.2.1 Main experiment

We start by eliciting baseline data on social media use. To fix the interpretation of the term "social media" and focus on apps that people find particularly tempting (Allcott et al., 2022), we inform all participants that we use this term to refer to the following apps and platforms throughout the survey: TikTok, Instagram, Snapchat, Facebook, YouTube, Reddit, and X/Twitter. We then randomly assign participants to take one of four different conversations with an AI chatbot in equal proportions, which participants complete in 14 minutes on average.[2] After completing the interview, we inform respondents that they will be asked some questions about their views on social media, emphasizing that there are no right or wrong answers to mitigate concerns about experimenter demand and social desirability bias (Bursztyn et al., 2025, forthcoming). We then elicit our main outcomes of interest. We provide a brief overview of outcomes here and introduce the outcomes in more detail when presenting the results from our experiment below.

We first elicit our two primary outcomes, as outlined in our pre-specification. Our first primary outcome is participants' motivation to reduce their daily social media use on an 11-point scale from 0 (Not at all motivated) to 10 (Extremely motivated). Our second main outcome is the perceived costs and benefits of social media. We elicit this using a 4-item battery of Likert scale questions that we combine into an overall index. We subsequently collect data on our secondary outcomes, including awareness of self-control problems in the context of social media (3-item battery), perceived self-efficacy (3-item battery), perceived impact of social media on one's quality of life (11-point scale), predicted future social media use, and participants' ideal amount of time spent on social media.

To evaluate how participants' perception of the conversations differ across treatment arms and the quality of the interviews according to established MI standards, we use the validated Client Evaluation of Counseling Scale that asks participants to describe the behavior of the

[1] Figure B.2 shows the distribution of social media use in our sample.

[2] Figure B.3 shows the distribution of interview duration by treatment arm.

interviewer (Madson et al., 2015). We additionally elicit emotional states during the interview using a subset of the PANAS scale (Watson et al., 1988).

We next elicit participants' willingness to pay (WTP) for a one-year premium access to a screen time management app that provides users with fine-grained control over their social media app use. We elicit willingness to pay using an incentive compatible multiple price list where participants choose between one-year access to the app and increasing amounts of bonus payments.[3] We then inform participants about the implied bounds on their WTP derived from their responses to the multiple price list and provide them with the opportunity to potentially revise their initial responses, with 20% revising their WTP downwards and 2% revising their WTP upwards. We use the revised responses to construct our final WTP measure. Importantly, we incentivize this choice to alleviate concerns about experimenter demand and social desirability bias for this measure (Bursztyn et al., 2025, forthcoming). Participants are informed that we will implement the decisions of 100 randomly selected participants. The survey ends with a battery of sociodemographic questions and an open-ended survey question about perceived study purpose.

2.2.2 Follow-up survey

To assess whether treatment effects persist, we conducted a follow-up survey more than two weeks after the main experiment. The recontact rate was high at 84.5%, and Appendix Table A.2 shows no evidence of selective attrition across treatment arms. Table A.3 presents summary statistics for the follow-up survey.

We re-elicited participants' motivation to reduce social media use, their perceived benefits and costs of social media, and whether they think social media makes their lives better or worse. We also collected self-reported social media time over the preceding two weeks. We next measured whether and how participants had attempted to reduce their usage. Furthermore, iPhone users with Screen Time enabled were asked to upload screenshots of their social media time, as recorded by the *Social* category for the preceding two weeks, allowing us to examine whether self-reports are a good proxy for verified use.

2.3 Treatments

All participants engage in a conversation with an AI chatbot (Figure B.4 provides a screenshot of the interface). We vary the nature of these conversations across treatment arms by instructing the AI interviewer with different conversation protocols and guidelines for the interviews. The conversations last around 14 minutes (as shown in Figure B.3). All conversation protocols share a common structure: they begin by exploring participants' views on social media, proceed to questions about their motivation to change their usage and confidence in their ability to do so, and

[3]The multiple price list includes the following values: \$0, \$1, . . . , \$10, \$15, \$20. We use \$0 and \$22.50 as the WTP for respondents that always prefer the bonus or the app, respectively. Otherwise, we use the midpoint of the implied WTP bounds.

conclude with a planning stage focused on reducing social media use. This consistent structure allows us to systematically vary (i) the conversational focus and (ii) the interviewer behavior while holding constant the medium, duration, and degree of participant engagement. Appendix Section D shows one exemplary interview for every condition and Table 1 provides an overview of the structural differences across treatment arms, which we discuss in more detail below.

2.3.1 Change Talk

The AI interviewer in this condition follows standard motivational interviewing (MI) practices: using open-ended questions to help participants discover their own motivation for reducing social media use, rather than telling them what to do or why. The chatbot employs MI techniques such as reflective summaries, active listening, and expressions of empathy to maintain a collaborative atmosphere.

The defining feature of this condition is its emphasis on eliciting pro-change arguments ("change talk") while minimizing the role of arguments in favor of the status quo ("sustain talk"). The focus on evoking change talk while responding to—but not amplifying—sustain talk is in line with current best practice for MI interviews (Miller and Rose, 2015).

The conversation opens by asking participants to share "the first thing that comes to mind when you think about your social media habits," followed by an exploration of how social media negatively affects them and conflicts with their personal values. This discussion deliberately focuses on negative aspects of excessive social media use to evoke change talk and reduce ambivalence.

The AI interviewer then introduces a scaling question, a standard MI tool: "On a scale from 0 to 10, how important is it for you to reduce your social media use?" Most participants respond between 5 and 7 (see Figure B.5). The key follow-up question—"Why did you say [X] and not a lower number like 0?"—prompts participants to generate their own reasons for reducing social media use.

This approach is then repeated to build confidence: We first elicit confidence to reduce social media use on a scale of 0-10 and then ask the follow-up: "Why did you say [X] and not a lower number like 0?" To further reinforce change talk, participants are subsequently asked to identify past examples of successful behavioral change.

The interview concludes with a planning phase, standard in MI, where participants identify small, concrete steps they could take. The final question prompts the interviewee to reflect on their key takeaway from the conversation.

2.3.2 Decisional Balance

While current MI best practice focuses on minimizing ambivalence by focusing on the advantages of change (Miller and Rose, 2015), early formulations of MI considered "working with

Table 1: Interview structures across treatment arms

T1: Change Talk	**T2: Decisional Balance**	**T3: Persuasion**
Discovering one's own motivation • What comes to mind when thinking about your social media? • What is negative about social media? • Follow-up about negatives • Conflict with personal values • Summary of discussion	*Discovering one's own motivation* • What comes to mind when thinking about your social media? • What is positive about social media? • Follow-up about positives • What is negative about social media? • Follow-up about negatives • Summary of discussion	*Information provision* • Overview of scientific evidence showing negative effects of social media use. Does this fit with your own experience? • Where would reducing social media make the biggest change for you?
Importance • How important is reducing social media use to you? • Why is it not a lower number? • If you were to reduce, what would be the most important change for you?	*Importance* • How important is reducing social media use to you? • Why is it not a lower number? • Why is it not a higher number? • If you were to reduce, what would be the most important change for you, good and bad?	*Negative aspects of social media* • Which apps are the biggest time sinks? • What is the biggest cost of using social media for you? • What benefits would you miss? • Follow-up moving conversation back to negatives
Confidence • How confident are you that you can reduce your social media use? • Why is it not a lower number? • Which personal strengths would help you achieve your goals? • When did you successfully manage to change a behavior in the past?	*Confidence* • How confident are you that you can reduce your social media use? • Why is it not a lower number? • Why is it not a higher number? • When did you successfully manage to change a behavior in the past?	*Choosing a plan* • How important is reducing social media use to you? • What is a concrete plan that you could implement now? • Offer two options of how to reduce social media use. Ask participants to choose. • Ask participants to restate their plan as a clear rule that they can commit to.
Planning & summary • What are small and realistic things you could do to reduce your social media use? • Ask participant to make it more specific and actionable • What is the most important takeaway from this conversation? • Summary of the key points	*Planning & summary* • What are small and realistic things you could do to reduce your social media use? • Ask participant to make it more specific and actionable • What is the most important takeaway from this conversation? • Summary of the key points	*Implementation & summary* • How will you enforce this rule going forward? • How confident are you that you will follow through? • What could you do to feel more confident? • Suggest actions to overcome barriers. What will you do now? • Summary of the key points
General interview style: • MI-consistent behavior • Redirects sustain talk towards change talk • No active attempt at inducing ambivalence about behavioral change • Reflections and summaries are biased towards participants' change-oriented statements	**General interview style:** • MI-consistent behavior • No active attempt to mitigate sustain talk • Explores both pros and cons of behavioral change (decisional-balance) • Two-sided reflections and summaries that give equal weight to both positives and negatives	**General interview style:** • Asks leading, closed-ended questions emphasizing need to change • Actively counter-argue • Actively push the participant to committing to reducing social media use • Reflections and summaries include comments and judgement from the interviewer

Note: This table provides a general overview of the structure of the conversations in our three treatment arms. Each column describes the nature and sequence of questions and summarizes the interviewing style of the chatbot.

ambivalence" through counterbalancing the pros and cons of change as the heart of MI (Miller et al., 1993). There is still an ongoing discussion about the relative merits of a change-focused strategy and a "decisional balance" (DB) approach that actively explores both pros and cons of change.

To study the role of ambivalence, we designed our second treatment to be largely identical to *Change Talk* but with a stronger emphasis on exploring ambivalent feelings by exploring both the pros and cons of change in a balanced way, consistent with best practice for MIs that include DB components. To achieve this, we implement three major changes compared to our first treatment. First, at the beginning of the interview, the AI chatbot explicitly asks participants to articulate both positive and negative effects of social media use, exploring conflict without attempting to resolve it or steer the conversation toward change. Second, the scaling questions on importance and confidence are each followed by two questions: why participants did not state a *lower* number and why they did not state a *higher* number. This symmetry ensures that participants actively explore their ambivalence about both the desirability and feasibility of change. Third, when summarizing participants' statements, the AI chatbot gives equal weight to pros and cons and does not redirect the conversation toward change talk.

2.3.3 Persuasion

The third condition departs sharply from MI principles by taking an active stance that the respondent *should* change their behavior and by giving unsolicited advice and information. For this purpose, the AI interviewer begins by presenting scientific evidence on the negative effects of social media use. This first question included an unsolicited information provision and was standardized to make sure the evidence was grounded in actual scientific evidence (Allcott, 2013; Braghieri et al., 2022; Haidt, 2024):

> I would like to have a conversation with you about your social media habits.
>
> To set the stage, here is a quick summary of what research tends to find: in several large studies, people who took a short break from social media or cut back sharply for a few weeks often reported feeling better overall and spending more time with friends and family offline. Other research also finds that very heavy daily use is linked with worse sleep and feeling more down or stressed.
>
> Taken together, this suggests many people are better off using social media much less than they currently do, especially if it has been affecting their focus, sleep, or time with others.
>
> How does this land with you—does any of it fit your own experience with social media?

The AI interviewer then continues to ask leading questions favoring change, actively counterarguing when participants express resistance, and pushes participants to commit to concrete actions

for reducing their social media use. At the same time, the AI interviewer is instructed to treat respondents with respect, so the tone is very similar to the other treatments, while still taking a clear stance that respondents should strive to reduce their social media use.If the respondent points out benefits or defends their social media use, the interviewer is instructed to acknowledge and briefly validate their perspective but then offer a gentle counter-perspective by pointing out overlooked downsides or relevant research. Furthermore, the interviewer is given an overview of current research findings and is instructed to only provide information grounded in this evidence. Through these steps, this treatment provides a clean test of whether attempts at direct persuasion can be effective in behavioral change or backfire by triggering psychological resistance towards change.

2.3.4 Control condition

To account for the effects of engaging in an AI-led conversation per se, and to keep survey duration and overall experience comparable, the control group participates in an AI-led interview about how they generally spend their time. This interview focuses on participants' daily routines (morning, afternoon, and evening) and how these routines vary (e.g., weekdays vs. weekends). Crucially, the conversation includes no motivational or persuasive elements related to social media and does not ask participants about their social media use.

2.4 Discussion of design

While the treatments differ in their focus—change-focused self-persuasion, self-persuasion with decisional balance, and direct persuasion—they were designed to be similar in length and follow a roughly identical structure.[4]

The key differences lie in how the interviewer responds to participants. Both *Decisional Balance* and *Change Talk* follow best practices for motivational interviewing by remaining neutral and making the respondent generate their own arguments for change. In contrast, the *Direct Persuasion* interviewer takes the concrete position that the respondent should cut down on social media consumption. Throughout the interview, the interviewer comments that the participant may not yet be fully committed, notes that many people only recognize the benefits of reduced social media use after trying it, and uses this as a bridge to request a small, measurable action the participant would be willing to take.

The contrasts in style are especially clear when looking at responses to the first scaling question. Consider respondents from our sample who rate the importance of reducing social media at a 5 out of 10. In both MI treatments, the AI interviewer follows up to explore the participant's reasons and asks: "Why is it a 5 and not a lower number like zero?", which directly prompts the interviewee to reflect on why they consider a reduction worthwhile. However, only in the *Decisional Balance* treatment, the interviewer additionally asks: "Why is it a 5 and not a higher

[4] See Figure B.3 for interview duration by treatment arm.

number like 10?", prompting the interviewee to think about reasons why current social media consumption levels could be worth sustaining. By contrast, the *Direct Persuasion* interviewer actively interprets the answer "5" as evidence of readiness to change and immediately shifts to planning: "You rated the importance as a 5, which suggests you see some meaningful upside in cutting back. What specific, measurable cutback plan would you be willing to try for the next few weeks?"

2.5 Implementation of AI interviews

To conduct AI interviews at scale, we use the infrastructure developed by Chopra and Haaland (2023). We embed a chat interface into our Qualtrics survey that allows participants to interact with an AI chatbot. Participants can respond by typing or by recording voice messages that are automatically transcribed and then deleted. This flexibility allows participants to choose the mode that best suits them; approximately 16% recorded at least one audio message.[5] Figure B.4 presents a screenshot of the chat interface.

Every interview question is generated by using a prompt tailored to the treatment assignment and the current stage of the interview. A prompt consists of three parts: (1) general behavioral guidelines for the interviewer in every treatment (e.g., "Your task is to conduct a high-quality motivational interview focused on helping participants discover and strengthen their own motivation for reducing the time they spend on social media."); (2) the history of the interview; and (3) a part specific to the individual question (e.g., "Ask how well their current social media habits fit with the kind of life they want"). We use OpenAI's current frontier model `GPT-5.2` model to generate responses. In Section 3.2, we validate our implementation by documenting that the AI chatbot adheres to core principles of motivational interviewing with high fidelity in our two self-persuasion treatments designed to follow MI best practices.

3 Descriptive evidence

Before presenting causal evidence on the effects of different conversations, we offer descriptive evidence to establish the relevance of the setting and document alignment with treatment protocols and MI fidelity. Specifically, in Section 3.1, we document that the overall majority of participants in our sample report feeling addicted to social media and express a preference for lower consumption levels. In Section 3.2, we validate our treatment implementation by showing that the AI interviewer behaved systematically differently across treatment arms, consistent with our design objectives and MI best practices. Finally, in Section 3.3, we characterize how the nature of conversations differed across treatment arms.

[5] We programmatically block the ability to paste text into the chat interface to prevent participants from using AI tools to answer the interview questions.

3.1 The wedge between actual and ideal social media consumption

We start by documenting that our participants would prefer to spend less time on social media before our intervention, which is the implicit premise of our interviews. At baseline, we ask participants to estimate their actual social media use and elicit their ideal social media use.

Overall, 82% of participants report an actual social media usage which exceeds their ideal level. On average, participants spend 190 minutes per day on social media apps but express a preference for only 104 minutes. Figure B.6 presents a binscatter plot of actual social media use (in hours) against ideal social media use (in hours). As expected, we find that ideal social media use is strongly positively correlated with actual social media use ($\rho = 0.671$, $p < 0.001$). However, we find that even heavy social media users prefer lower levels of consumption. In fact, we document an increasingly positive wedge between actual and ideal social media use along the full distribution of actual social media use.

We also elicit an introspective judgment of how addicted participants feel towards the seven social media apps we focus on in our survey (Facebook, Instagram, Reddit, Snapchat, TikTok, X/Twitter, YouTube) on a 5-point scale from "Not at all addicted" to "Very addicted" (while also allowing for "I do not use this app"). 64% of participants report being at least "somewhat addicted" to one of these social media apps, with TikTok (44%), Instagram (30%) and YouTube (33%) being the apps with the highest self-reported share of addiction. These apps are also the most used in our sample, with an average daily screen time of 31 minutes per day for TikTok, 31 minutes per day on Instagram, and 43 minutes per day on YouTube. These descriptives underscore that the majority of our sample of social media users feels strongly addicted to social media apps and would prefer to reduce their social media use, in line with results from previous studies on social media (Allcott et al., 2022; Bursztyn et al., 2025).

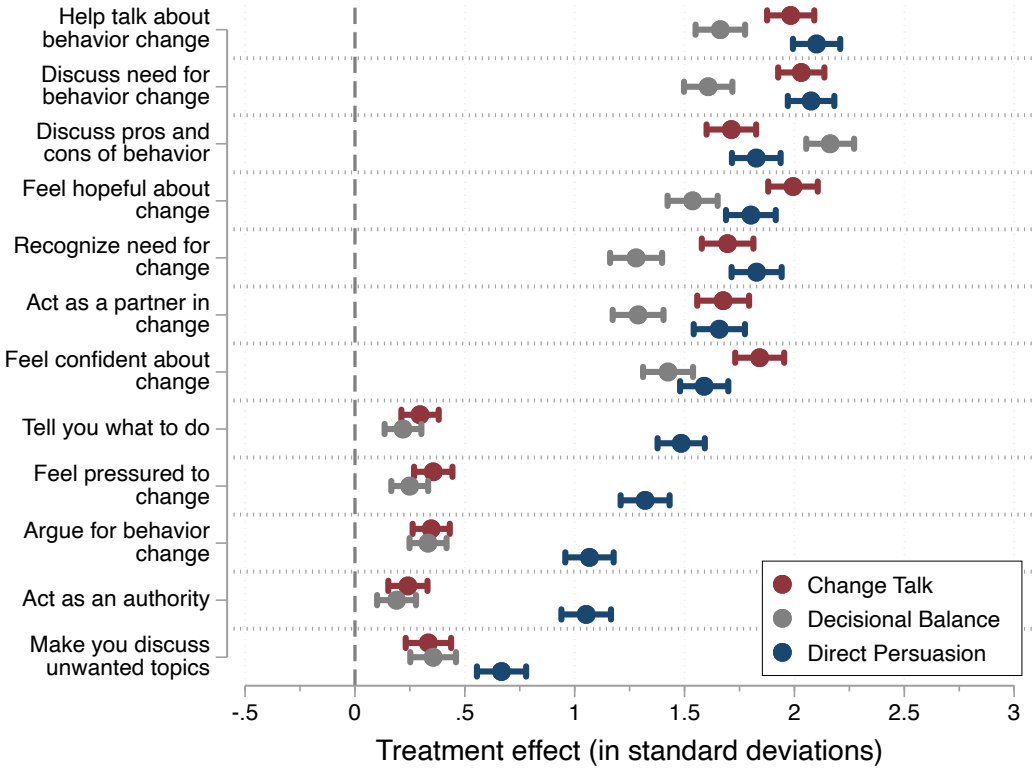
3.2 Validation of treatment implementation

We next provide evidence that we managed to change the nature of the conversations across treatment arms as intended. We first present evidence based on how participants rate the behavior of the AI chatbot across treatment arms. We then present evidence from a validated, LLM-based annotation pipeline demonstrating that our motivational interviews were delivered with high fidelity and adherence to best practices for motivational interviewing.

3.2.1 Participants' evaluations of interviewer behavior

To document that the AI-led conversations vary systematically across treatment arms, we elicited participants' evaluation of interviewer behavior using the validated Client Evaluation of Counseling Scale (CECS, Madson et al., 2013). Self-reported perceptions not only test whether conversations are technically different, but also whether these differences are salient enough to be noticed by participants.

Figure 1: Participants' self-evaluation of the interviews



Note: This figure shows treatment effect estimates from OLS regressions on perceptions of interviewer behavior, using the time use interview control group as the omitted category. We separately show treatment effect for all 12 items of the validated Client Evaluation of Counseling Scale. Each item is scored on a 5-point Likert scale from "not at all" to "always". We standardize each item to have a mean of zero and a standard deviation of one among control group respondents. Positive coefficients indicate that participants perceive the interviewer as displaying more of the corresponding behavior. 95% confidence intervals derived from robust standard errors are shown.

Figure 1 shows that participants' evaluations of interviewer behavior closely track the intended design of the conversational protocols. As a first sanity check, it is reassuring to see that compared to control group respondents—who did not have a conversation about behavioral change—participants in both the MI and the persuasion treatment arms perceived the interviewer as being about two standard deviations more focused on (i) helping them talk about behavioral change and (ii) discussing the need for behavior change.

Turning to the MI treatment arms, we find that *Change Talk* strongly increases perceptions of other MI-consistent behaviors: respondents are substantially more likely to report that the interviewer (i) helps them discuss the pros and cons of behavior change, (ii) makes them feel more hopeful about change, (iii) acts as a partner in change (rather than an authority), (iv) helps them recognize the need for change, and finally (v) makes them feel more confident about change. At the same time, *Change Talk* produces only small increases in perceptions of being told what to do or being pressured to change, indicating a predominantly autonomy-supportive interaction.

Evaluations of the *Decisional Balance* conversations are, as expected, close to the *Change Talk* evaluations in that participants similarly perceive the interviewer as systematically displaying MI-consistent behaviors. However, there are notable differences along the expected margins. First, *Decisional Balance* scores highest in terms of discussing *both* the pros and cons of behavior change, in line with the instructions to provide pros and cons symmetrically. As a consequence of this more balanced conversational approach, it is not surprising to observe that participants in the *Decisional Balance* treatment rate the interviewer lower on helping them (i) recognize the need for change, (ii) feeling hopeful or (iii) confident about change, compared to the *Change Talk* condition. This suggests that the AI interviewer in the *Change Talk* treatment was successful in the objective to evoke change, while the AI interviewer in the *Decisional Balance* treatment induced stronger feelings of ambivalence by focusing more on both the pros and cons of change, in line with the decisional balance approach.

For the *Direct Persuasion* treatment, we see that the interviewer—to a similar extent as the *Change Talk* treatment–led respondents to talk about behavior change, discuss the need for change, act as a partner in change, and other MI-consistent behaviors. However, compared to both the MI treatment arms, the *Direct Persuasion* generates a sharply different evaluation on MI-inconsistent behaviors. Compared to the MI treatments, participants in the *Direct Persuasion* treatment strongly perceive the interviewer as (i) arguing for behavior change, (ii) telling them what to do, (iii) acting as an authority, (iv) exerting pressure to change, and (v) discussing unwanted topics. This validates that the conversations in the MI treatments were indeed more collaborative and non-directive than the direct persuasion treatment.

Overall, the perception data strongly validates our treatment implementation: *Change Talk* is perceived as a change-focused motivational interview, *Decisional Balance* as an MI with decisional balance, and *Direct Persuasion*—while showing some MI-consistent behaviors, such as helping them talk about behavior change and recognize the need for change—also inducing many persuasive elements in direct conflicts with MI protocols, such as acting as an authority and making respondents feel more pressured to change.

3.2.2 Emotional states

Having established that participants perceived the conversational protocols as intended, we next examine whether these distinct interviewing styles also produced the hypothesized emotional responses. Figure B.7 shows that the three conversational protocols induce sharply distinct emotional profiles during the interview, closely mirroring their intended mechanisms. *Change Talk* generates a strongly action-oriented emotional state. It substantially increases feelings of determination and encouragement by about 0.9-1.1 standard deviations, while also raising interest and excitement. At the same time, it produces moderate increases in aversive self-conscious emotions such as guilt and shame, with comparatively small effects on irritation or being upset. This pattern is consistent with change-directed self-persuasion: participants come to recognize

costs and personal responsibility, but this recognition is paired with empowerment rather than defensiveness, indicating that the internally generated nature of the arguments largely limits psychological reactance.

Decisional Balance produces somewhat similar emotional profile, but the effects on determination and encouragement are significantly weaker compared to *Change Talk* ($p < 0.10$). By giving equal weight to reasons for and against change, the decisional balance protocol appears to engage participants cognitively without pushing them towards commitment. At the same time, it does not provoke strong resistance or reactance, leaving participants emotionally engaged but in a less mobilized state.

Direct Persuasion induces the most emotionally conflicted response. While it does increase feelings of determination and encouragement, these positive effects are weaker compared to *Change Talk* ($p < 0.05$). At the same time, persuasion produces notably stronger increases in irritation and feeling upset. Overall, the changes in perceptions and emotions associated with the *Direct Persuasion* treatments correspond to factors predicted by psychological reactance theory to produce resistance towards change (Steindl et al., 2015).

3.2.3 LLM-based measurement of the quality and fidelity of motivational interviews

Two related empirical challenges in evaluating the effects of MIs and aggregating evidence from separate MI trials are the heterogeneity in the quality and consistency of MI delivery and the frequent lack of documentation on adherence to protocols (Frost et al., 2018; Zhu et al., 2024). As a complement to participant's perception, we therefore follow the gold standard in the literature and assess the fidelity of MI implementation with the validated Motivational Interviewing Treatment Integrity coding manual (MITI 4.2.1, Moyers et al., 2014, 2016).

The MITI coding manual consists of four global scores: (i) cultivating change talk, (ii) softening sustain talk, (iii) partnership, and (iv) empathy.[6] The first two scores are often referred to as the *technical* component, while the latter two capture the *relational* component. Scores for individual components range from 1 (low) to 5 (high), and should be assigned based on a holistic evaluation of the full transcript. Importantly, the MITI scores should only evaluate the interviewer, meaning that an interviewee more reluctant to change does not lead to low scores.

We develop and validate a scalable LLM-based workflow for scoring interviews on the MITI inventory. Specifically, we design prompt templates for each of the four global scores. Each prompt contains (i) a short description of the task and role of the LLM as annotators of motivational interviews, (ii) the full verbatim instructions from the MITI coding manual for this score, (iii) the full transcript of the conversation under evaluation, and (iv) LLM-specific annotation instructions and output constraints. See Appendix Section C.1 for further details and the full prompt template.

[6] We refer the reader to the extensive MITI coding manual for definitions of the individual scores, as well as further information on how to annotate interviews: `https://casaa.unm.edu/assets/docs/miti4_21.pdf`

We validate the LLM-based measurement against a ground truth derived from human annotated motivational interviews. Specifically, we obtain the 14 annotated transcripts that are part of the official training materials to demonstrate both high and low quality implementations of MI principles encoded by the MITI 4.2.1 coding manual.[7] These interviews have been annotated by researchers working on motivational interviewing. In Appendix Section C.1, we show that LLM measurements closely align with human-assigned scores, achieving a correlation of 0.72 when pooling all four scores ($N = 56$). If anything, we find that LLMs are slightly more conservative in scoring MI transcripts with a mean bias of -0.24 on a 5-point scale.

Figure B.8 presents the results from annotating all 1,369 motivational interviews from the *Change Talk* and *Decisional Balance* arms with the LLM-validated workflow. As expected, we find that conversations in both treatment arms score very high on each of the four global measures from the MITI coding manual.[8] This suggests that we were able to design an AI chatbot that systematically adheres to best practices for MIs. Indeed, we find that scores are strongly concentrated, suggesting that delivering MIs with the help of AIs might help decrease interviewer variance in delivery, which has been singled out as an issue for scaling up MIs (Hallgren et al., 2018).

3.3 How do topics differ between conversations?

We have demonstrated that (i) participants perceive large and systematic differences across treatment arms in line with our design goals and that (ii) the delivery of our motivational interviews complies with best practices for MIs. In this section, we further examine how the conversations differ across treatments using a combination of unsupervised topic modeling and targeted quantitative metrics motivated by the results from the unsupervised analysis.[9]

3.3.1 Topic analysis

To obtain a comprehensive overview of the themes appearing across conversations, we conduct an unsupervised topic analysis. We define documents as participants' individual responses to questions from the AI interviewer, excluding the control group and very short responses with fewer than 10 characters. Our final corpus consists of 25,952 documents. We construct embeddings using the `all-mpnet-base-v2` sentence transformer, reduce dimensionality with UMAP, and apply BERTopic (Grootendorst, 2022) to identify the 30 most frequent topics. For interpretability, we prompt `gpt-4.1-mini` to generate a short label for each topic based on the 10 documents with the highest topic probability and merge four topics with high overlap. Finally, we construct binary indicators for whether each topic appears at least once in an interview transcript.

[7] Available at `https://casaa.unm.edu/tools/miti.html`

[8] The MITI manual suggests a threshold of 3.5 for "fair" and a threshold of 4.0 for "good" MI proficiency, the latter being the highest category.

[9] In Appendix Section C.2, we present supporting evidence on how people responded to the scaling questions.

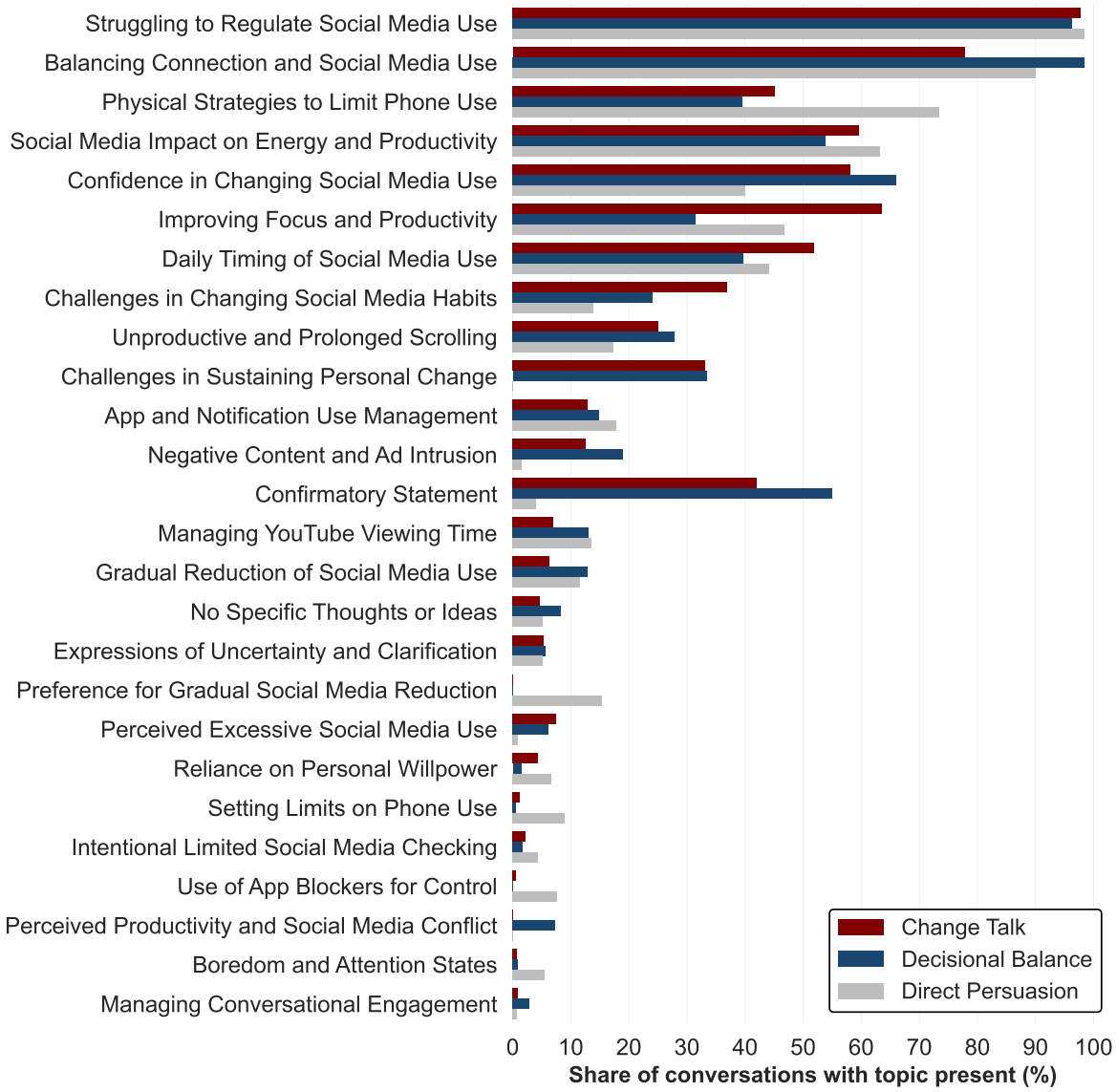
Figure 2 presents the frequency of each topic separately by treatment arm. Several patterns emerge that align with the design of our interventions. First, nearly all conversations across conditions discuss struggling to regulate social media use, confirming that participants engaged meaningfully with the core subject matter. However, the thematic emphasis diverges substantially across treatment arms. Second, participants in *Decisional Balance* are approximately 20 percentage points more likely to discuss the trade-off between benefits and costs of social media ("Balancing Connection and Social Media Use") compared to *Change Talk* (95% vs. 75%). This pattern is consistent with the *Decisional Balance* protocol's emphasis on exploring both positive and negative aspects of social media use. Third, conversations in *Direct Persuasion* focus considerably more on concrete action and implementation strategies. Topics such as "Physical Strategies to Limit Phone Use" and "App and Notification Use Management" appear more frequently in this condition, reflecting the directive nature of the intervention, which emphasizes telling participants what to do rather than eliciting their own motivations. Fourth, participants in the MI-based interviews (*Change Talk* and *Decisional Balance*) are substantially more likely to discuss their confidence in making behavioral changes and the challenges of sustaining personal change. This aligns with MI's core technique of eliciting self-efficacy statements and exploring barriers to change. Notably, *Change Talk* participants also show a distinct spike in expressing a "Preference for Gradual Social Media Reduction," consistent with the protocol's focus on eliciting participants' own commitment to change.

3.3.2 Metrics

Motivated by the results from the above topic analysis, we aim to more directly quantify three aspects of the interview content: First, how many distinct positive aspects of social media are mentioned by the interviewee? Second, how many distinct negative aspects? And third, how often do different strategies for reducing social media show up in conversations? We extract these metrics by prompting `gpt-5-nano` to identify and count these elements in each transcript.

Positive versus negative aspects Figure B.9 shows that participants in *Change Talk* and *Direct Persuasion* identify three positive aspects of social media use on average. In contrast, participants in *Decisional Balance* mention 50% more positive aspects. Similarly, we find *Decisional Balance* produces the largest number of negative associations with social media use, followed by *Change Talk* and *Direct Persuasion*. This underscores that conversations in *Decisional Balance* are more focused on a systematic exploration of both the pros and cons of social media use, while *Change Talk* is more focused on making participants talk about the negative aspects that would motivate them to change. The finding that *Direct Persuasion* produces the least amount of positive and negative aspects overall reflects the stronger focus of this conversation on making people commit to concrete actions and strategies for reducing social media use, to which we turn next.

Figure 2: Unsupervised topic analysis of conversations



Note: This figure presents the results from an unsupervised topic analysis of our interview transcripts with BERTopic. For each treatment arm, we present the share (in %) of conversations with a given topic appearing at least once during the full interview. See the main text for details.

Strategies Figure B.10 displays how frequently participants mention different strategies for reducing social media consumption during interviews.[10] The three most commonly proposed strategies across conditions are setting rules or goals, reducing overall phone use (rather than targeting social media specifically), and relying on willpower and discipline. The most striking pattern is that *Direct Persuasion* explicitly elicits discussion of concrete rules and goals, ensuring these topics appear in virtually all interviews. In the MI treatment arms, such structured

[10]We constructed a coding manual of different strategies after a systematic review of transcripts. We construct indicator variables for having discussed a given strategy by prompting `gpt-5-nano` to identify whether the strategy was mentioned during the conversation.

approaches are less often discussed, as the interviewee must generate their own strategy for reduction without such explicit prompts.

The analysis of interview content reveals clear differentiation across treatment arms in both substance and focus. Whether these distinct conversational experiences translate into differential effects on motivation, beliefs, and actual social media consumption is the question we address next.

4 Causal effects of conversations

In this section, we examine the causal effects of different conversational protocols. We begin by outlining our empirical specification in Section 4.1 and then present treatment effects on motivation to change social media use and related beliefs in Section 4.2. We next turn to an incentivized measure of willingness to pay for a screen time app in Section 4.3, and conclude with evidence on actual social media time use in Section 4.4

4.1 Empirical specification

We follow our pre-specification and analyze the effects of our treatments by estimating the following primary specification with ordinary least squares:

$$Y_i = \alpha + \beta \text{Change Talk}_i + \gamma \text{Decisional Balance}_i + \delta \text{Direct Persuasion}_i + \xi X_i + \zeta Y_i^0 + \varepsilon_i \quad (1)$$

where Y_i denotes the outcome of interest. The main independent variables are indicator variables for treatment assignment, *i.e.*, $\text{Change Talk}_\text{i}$ equals one for participants assigned to the motivational interview focused on inducing change talk, and zero otherwise. $\text{Decisional Balance}_\text{i}$ equals one for participants assigned to the motivational interview aimed at decisional balance, and zero otherwise. $\text{Direct Persuasion}_\text{i}$ equals one for participants assigned to the direct-persuasion interview, and zero otherwise. The omitted category is the AI interview about time use. We note that the coefficients on the treatment indicators should be interpreted as intention-to-treat effects. We include a standard set of preregistered control variables (X_i) and, if available, control for pre-treatment values of the outcome of interest (Y_i^0).[11] We use robust standard errors clustered at the participant level for inference.

[11] The control variables are age, gender (female indicator), log household income, having at least a college degree (binary indicator), full-time employment (binary indicator), above-median baseline average daily social media use (binary indicator), and separate indicators for spending at least five minutes on average per day on TikTok, Instagram, Snapchat, YouTube, Facebook, and X/Twitter. For the follow-up survey, we preregistered to control for baseline social media time continuously.

4.2 Motivation and beliefs

4.2.1 Motivation to change

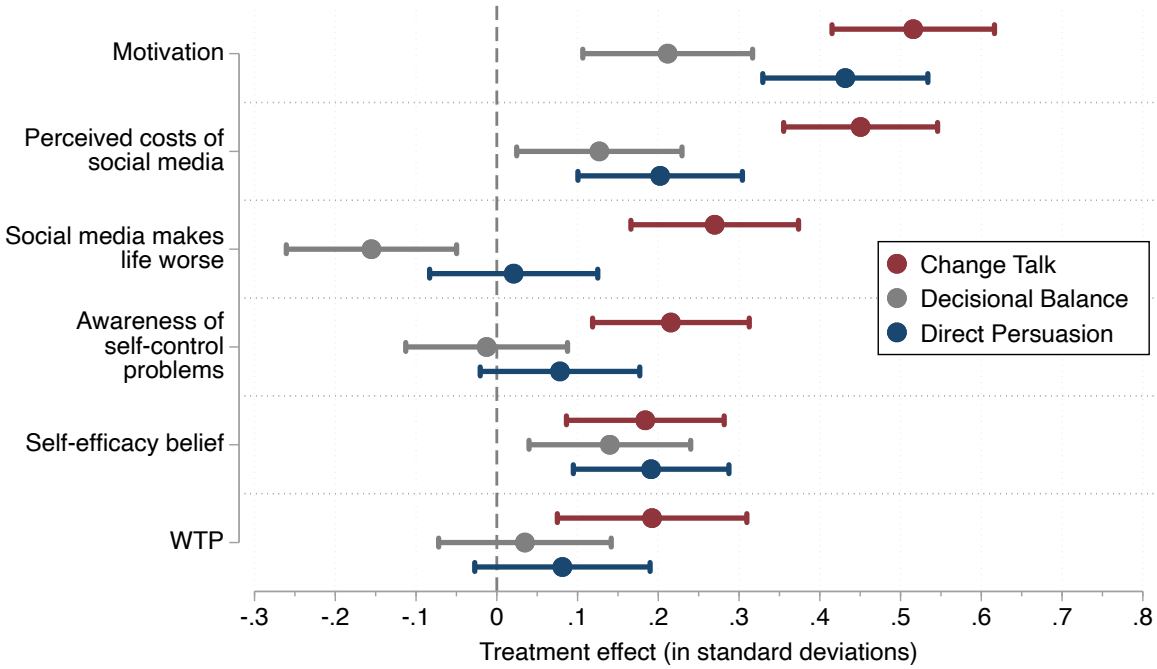
Our first main outcome of interest is people's motivation to reduce their social media time after the conversation, as it directly maps with our focus on motivational interviews. Furthermore, inducing a state of mind where people are willing to take a first step is a key bottleneck of behavioral change (Deci and Ryan, 1985). Figure 3 presents treatment effects estimates visually (Appendix Table A.4 provides the corresponding regressions in table form). *Change Talk* yields the largest effects with an increase of 0.52 standard deviations, followed by *Direct Persuasion* with 0.43 standard deviations. While the difference between these treatment arms is not statistically significant at conventional levels ($p = 0.101$), both treatments are statistically significantly more effective than *Decisional Balance*, which yields a smaller but still substantial effect size of 0.21 standard deviations.

To benchmark these effect sizes, we turn to the control group. Specifically, control group participants with above-median baseline social media use report a 0.24 standard deviation higher motivation to reduce their social media time than participants with below-median baseline social media use, with a mean difference in social media use of 186 minutes per day between the two groups. This natural comparison underscores that our conversations generate economically meaningful effect sizes on people's motivation to reduce social media use.

Facets of motivation To gain a more nuanced understanding of the motivational effects, we complement our main measurement of motivation to change with a more fine-grained measure of motivation. Specifically, we elicit six progressive facets of motivation to change with items from the Change Questionnaire that are grounded in psycholinguistic analysis of natural language (Amrhein et al., 2003). We elicit agreement with statements of the form "I [*verb*] to reduce my social media use" on a 5-point scale, where *verb* is substituted with "want", "could", "have good reasons", "intend" and "am trying to". These items thus elicit the strength of participants' (i) desire, (ii) ability, (iii) reasons, (iv) need, (v) commitment and (vi) plans to reduce their social media use.

Figure B.11 presents separate treatment effect estimates on each facet. Consistent with our main measurement of motivation, we replicate the qualitative ranking of treatment effect point estimates for each facet of motivation: *Change Talk* $>$ *Direct Persuasion* $>$ *Decisional Balance*. For example, *Change Talk* increases desire to change ("want") by 41% of a standard deviation, with a substantially weaker treatment effect of 16.9% of a standard deviation for *Direct Persuasion* and no change among participants in *Decisional Balance*. For three out of six facets (desire, reasons, intend), we estimate statistically significantly larger treatment effects for *Change Talk* compared to *Direct Persuasion* at the 5% level. *Direct Persuasion* generates statistically significantly larger treatment effects than *Decisional Balance* on each facet except

Figure 3: Treatment effects on motivation, beliefs and willingness-to-pay



Note: This figure plots treatment effect estimates of different conversational protocols obtained from specification 1. We standardize all dependent variables to have a mean of zero and a standard deviation of one in the control group. "Motivation" is participants' motivation to reduce their social media time. "Perceived costs of social media" is an index capturing the perceived costs and benefits of social media, with larger values indicating higher perceived costs of social media use. "Social media makes life worse" captures whether participants view their own social media consumption as making their life worse (rather than better), which is measured on an 11-point scale. "Awareness of self-control problems" is an index with higher values indicating greater awareness of behavioral self-control problems. "Self-efficacy beliefs" is oriented such that larger values indicate higher perceived self-efficacy. "WTP ($)" captures participants' standardized willingness to pay for an app that helps users to reduce their social media use by limiting time spent on social media apps. 95% confidence intervals derived from robust standard errors are shown.

reasons to change. Interestingly, *Decisional Balance* has no statistically significant effect on "desire", "ability" or "need" to change, which likely reflects the more ambiguous stance towards social media induced by a more balanced discussion of both the pros and cons of social media use (see Section 3.2 and 3.3 for further evidence).

Persistence We document the persistence of effects on motivation over a two-week period in our follow-up survey. Figure B.12 shows that while treatment effects are naturally attenuated, participants in *Change Talk* and *Direct Persuasion* still report 0.14 and 0.16 standard deviations higher motivation to reduce social media use compared to the control group ($p < 0.001$). The effects of *Decisional Balance* are more muted (0.07 SD). This suggests that a one-time 15-minute conversation can have sustained effects on people's motivation to change over a 2 week period.

4.2.2 Cost-benefit perceptions

Our second main outcome of interest are beliefs about the costs and benefits of social media use. There is ample empirical support for belief-based models of persuasion, in which persuasion works by changing the beliefs of the receivers (DellaVigna and Gentzkow, 2010). Our *Direct Persuasion* treatment tried to achieve this by providing direct information about the benefit of reduced social media consumption. Furthermore, it is possible that our MI treatments also affect beliefs through a self-persuasion channel or, in the case of the *Change Talk* treatment, selective attention on the positive aspects of change.

To elicit cost-benefit perceptions, we include a 4-item battery and construct a standardized index oriented such that larger values correspond to higher perceived costs of social media use. Figure 3 documents statistically significant effects for all treatments in the direction of higher perceived costs of social media use. *Change Talk* yields the largest effect of 0.45 standard deviation ($p < 0.001$), which is three times larger than the effect of *Decisional Balance* which produces a more modest treatment effect of 0.13 standard deviations ($p < 0.05$). This result is consistent with a selective attention story in which *Change Talk* focuses more on the negative aspects of social media use while *Decisional Balance* treats the positive and negative aspects in a more balanced way.

Turning to *Direct Persuasion*, we see a treatment effect of 0.20 standard deviations ($p < 0.01$). This effect size is significantly smaller than for *Change Talk* ($p < 0.001$), but is not statistically significantly different from *Decisional Balance* at conventional levels ($p = 0.158$). While both *Change Talk* and *Direct Persuasion* focused the conversation around negative aspects of social media use, *Direct Persuasion* directly confronted participants with negative consequences while *Change Talk* asks participants to identify negative aspects themselves. The substantially weaker treatment effect of *Direct Persuasion*—despite the provision of information about the negative consequences of social media use—is consistent with externally provided arguments being less effective than internally generated ones.

Figure 3 documents that the effects of *Change Talk* and *Direct Persuasion* on cost-benefit perceptions persist in our two-week follow-up survey ($p < 0.001$). The effect of *Decisional Balance* is no longer statistically significant in the follow-up survey, providing further evidence that striving for decisional balance in conversations aimed at facilitating behavioral change may be ineffective at changing the calculus of costs and benefits.

As a direct complement to the cost-benefit beliefs index described above, we also asked our respondents to make an overall evaluation of whether their *current* social media consumption "makes [their] life better or worse." We use an 11-point Likert scale from "-5 (makes my life worse)" to "+5 (makes my life better)." This measure thus endogenizes the evaluation of the cost-benefit trade-off at participants' own current usage level.

Figure 3 reveals pronounced differences between treatment arms for this measure. *Change*

Talk is the only conversational protocol that negatively affects participants' evaluation of social media ($p < 0.001$). In contrast, we document more positive views of social media in *Decisional Balance* compared to the time use control group, which likely reflects the fact that participants in *Decisional Balance* were prompted to also identify positive effects of social media use on their life. *Direct Persuasion*, by contrast, has no statistically significant effect. In our two-week follow-up survey, we observe qualitatively and directionally similar patterns (see Figure B.12), with a 0.09 standard deviation more negative evaluation of social media in *Change Talk* and a positive point estimate for *Decisional Balance*. However, treatment effects are more noisily estimated given the lower sample size and are not statistically significant at conventional levels.

4.2.3 Awareness and self-efficacy

In addition to our primary outcomes on motivation and cost-benefit beliefs, we also elicited measures of awareness of self-efficacy to further illuminate mechanisms behind our treatment effects.

Awareness Awareness of a behavioral self-control problem is an important step in recognizing the need for change (Prochaska and DiClemente, 1983). To elicit awareness about self-control problems in the domain of social media use, we designed a 3-item battery that elicits agreement with statements such as "Even when I try to limit my social media use, I find it difficult to stop", from which we construct a standardized index using the mean and standard deviation in the control group.

In Section 3.1, we documented that our sample of heavy social media users largely reports feeling addicted to social media apps. While this suggested a limited scope for further increasing awareness of self-control problems, an extended conversation about one's social media use could still change the way participants evaluate their past self. Indeed, Figure 3 shows that only *Change Talk* causes a significant increase in awareness of 0.22 standard deviations ($p < 0.001$), while the effects of *Direct Persuasion* and *Decisional Balance* are both statistically insignificant and economically smaller at less than 0.10 standard deviations. These patterns are consistent with memory models of selective recall (Bordalo et al., 2025), where prompting participants to engage in one-sided introspection of past behavior affects their beliefs (Conlon, 2025).

Self-efficacy A large literature in economics and psychology recognizes the importance of self-efficacy as a motivator for change (Bandura, 1977). Attempting to reduce social media use requires believing that such change is possible. The conversational protocols in our treatment arms include dedicated discussions of people's confidence in their ability to change, such as the scaling questions related to confidence or introspective questions about past examples of successful behavior change. It is therefore plausible that these conversations could affect self-efficacy.

In our main survey, we therefore elicit self-efficacy with a 4-item battery that includes

statements such as "When I set my mind to something, I usually succeed." To set a higher bar for detecting treatment effects, we do not refer to social media specifically, but focus on generalized self-efficacy beliefs. We construct an index standardized to have a mean of zero and a standard deviation of one in the control group.

Figure 3 documents statistically significant and comparatively uniform treatment effects of our interventions on self-efficacy beliefs that range from 0.14 standard deviations (for *Decisional Balance*) to 0.19 standard deviations (for *Direct Persuasion*). We cannot reject the null hypothesis of identical treatment effects (all $p > 0.285$). This suggests that conversations about the possibility of change in one domain can affect people's overall evaluation of their self-efficacy.

4.3 Willingness to pay

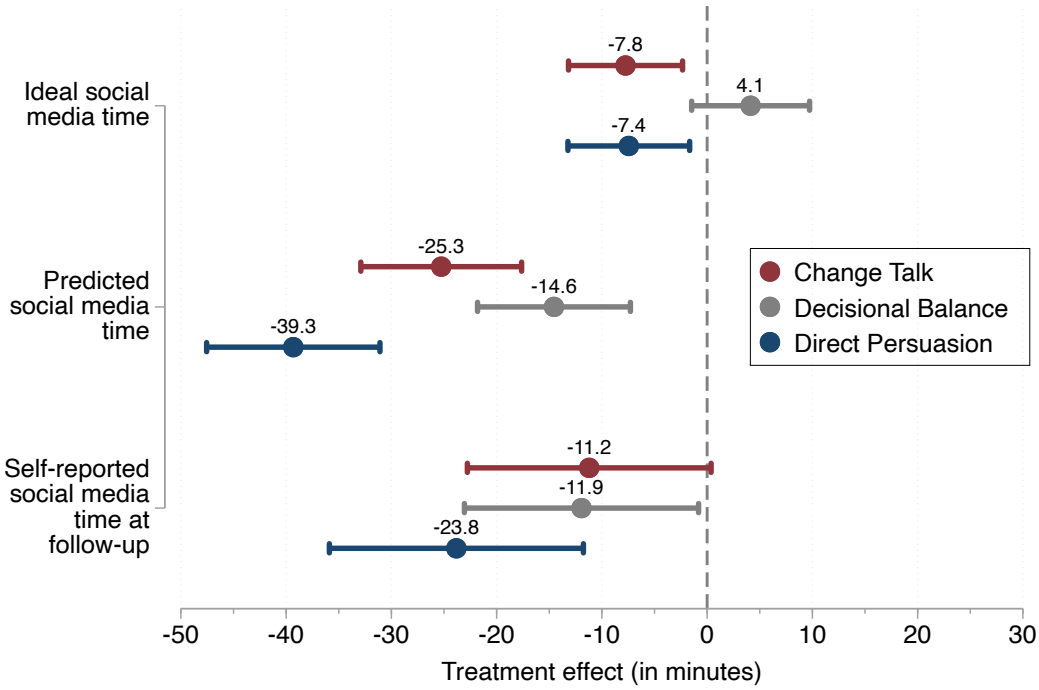
Having documented systematic and persistent changes in people's motivation in the previous section, we turn to a natural next question: Do our conversational protocols affect people's willingness to incur costs to reduce their social media use?

To address this question, we elicit participants' willingness to pay (WTP) for one-year access to a screen time management app that allows users to control their time spent on different apps across all of their devices, the *Freedom app*. We elicit willingness to pay in an incentive compatible way (see Section 2 for details), making this a costly action. This WTP measure helps alleviate concerns about experimenter demand effects that may arise with self-reported outcomes.

Table A.4 reports treatment effects on WTP. Consistent with *Change Talk* having the strongest effects on motivation and beliefs, we document a sizeable increase in WTP by \$0.82 ($p < 0.01$). This corresponds to an increase of 27% relative to the mean of \$3.00 in the control group. In contrast, neither *Decisional Balance* (\$0.15) nor *Direct Persuasion* (\$0.35) yields statistically significant effects on WTP. The consistency between the large and persistent self-reported motivational effects of *Change Talk* and its sizeable impact on incentivized WTP therefore strengthens the interpretation that this protocol generates genuine, behaviorally relevant increases in motivation rather than purely expressive responses.

Furthermore, the null effects on WTP for the other treatments do not imply that the self-reported increases in motivation associated with these treatments reflect experimenter demand effects (see Section 4.4.5 for a discussion about demand effects). Our WTP measure reflects demand for a very specific screen time strategy: strict app limits enforced by a third-party app. This strategy might not appeal to everyone even when they are motivated to reduce their social media use—instead preferring other strategies for regulating social media use—to which we turn in Section 4.4.4.

Figure 4: Treatment effects on ideal, predicted and actual social media time



Note: This figure reports treatment effect estimates using data from our main survey. Ideal social media time and predicted social media time are elicited in average minutes per day during the main survey. Actual social media time is self-reported in the two-week follow-up survey. We standardize all outcomes at the 5th and 95th percentiles. 95% confidence intervals derived from robust standard errors are shown.

4.4 Social media time use

Our earlier results show that the conversational protocols increase motivation for change and beliefs about the costs of social media use. In this section, we examine whether these shifts also translate into intended and actual behavioral change. We analyze the effects of our conversations on (i) participants' ideal amount of time spent on social media, (ii) their predicted social media use, and (iii) their actual self-reported social media use over the weeks following our main survey. Finally, we validate our self-reported measures using screenshots of social media time use from a subset of respondents.

4.4.1 Ideal and predicted time spent on social media

We distinguish between ideal social media time, which reflects participants' normative preferences, and predicted social media time, which reflects beliefs about one's future behavior. Our conversational protocols could arguably affect both margins, as all treatments affected participants' evaluation of the costs and benefits of social media use as well as perceptions of self-efficacy, making them more confident in their ability to reduce their social media use.

Figure 4 documents that our treatments indeed affect ideal and predicted social media time

(see Table A.5 for the underlying regression estimates). Turning first to ideal social media time, we find participants in *Change Talk* and *Direct Persuasion* report that they would like to spend between 7.4 and 7.8 fewer minutes per day on social media compared to the control group. This corresponds to an 8% and 9% decrease compared to the control group mean of 90 minutes per day. However as shown in Section 3.1, almost all participants already report substantially lower ideal social media time at baseline, suggesting that beliefs about the ideal social media time per day are not the primary barrier to change.

Turning to participants' predictions about their future social media use, Figure 4 reveals economically sizeable effects. Participants in *Change Talk* predict reducing their social media use by 25.3 minutes per day on average over the next two weeks ($p < 0.001$), while *Decisional Balance* yields a reduction of 14.6 minutes per day ($p < 0.001$). These effects correspond to a decrease of 16% and 9% compared to the control group mean, respectively. *Direct Persuasion* generates the largest treatment effect with a predicted reduction of 39.3 minutes ($p < 0.001$), which could reflect the stronger emphasis of this conversational protocol on discussions of strategies to turn intentions into actions.

App-level effects To better understand the sources of the overall effects on predicted time use, we next examine app-level predictions of future use. This analysis sheds light on which types of social media use are most responsive to the interventions. We elicit app-level predictions of changes in time spent on a 5-point scale from "much less" to "much more." For this analysis, we standardize responses to have a mean of zero and a standard deviation of one in the control group, after excluding participants who report not using a given app.

Appendix Figure B.13 reports treatment effects on predicted changes in time spent on seven major social media platforms. Across most apps, we observe economically meaningful reductions in planned usage, with point estimates typically on the order of 0.3–0.6 standard deviations relative to the control group. Snapchat is the only platform for which we do not detect systematic treatment effects, likely reflecting that it partly functions as an instant messaging app.

Consistent with the earlier results on predicted time use, the largest point estimates typically arise under *Direct Persuasion*. While *Change Talk* produces effects of similar magnitude for several apps, *Direct Persuasion* tends to generate the most pronounced predicted reductions in use, particularly for TikTok, Instagram, Facebook, YouTube, and Reddit. Although differences between *Change Talk* and *Direct Persuasion* are not statistically distinguishable for any individual app (all $p > 0.10$), the ordering of point estimates is remarkably stable across platforms. By contrast, the *Decisional Balance* treatment yields substantially more muted effects on intended app use, with point estimates that are typically around half the size of those observed under *Direct Persuasion* and *Change Talk*.

4.4.2 Actual time spent on social media

A central question is whether the effects on predicted changes in time use documented above translate into actual reductions in social media use. We examine this using self-reported time use from our follow-up survey conducted two weeks after the intervention. We then show that these self-reported data show a high correlation with verified social media screen time from screenshots that we obtained for a subset of respondents who use iPhones with Screen Time enabled, underscoring the reliability of the self-reported data.

Self-reported time use Figure 4 presents treatment effects on self-reported daily social media use in the follow-up survey, measured in average minutes per day over the previous two weeks for the seven apps we focus on (Facebook, Instagram, Reddit, Snapchat, TikTok, X/Twitter, YouTube). All three treatment arms show reductions in social media use relative to the control group. Participants in the *Change Talk* condition report social media reductions of 11.2 minutes per day ($p < 0.10$), those in the *Decisional Balance* condition report reductions of 11.9 minutes ($p < 0.05$), and those in the *Direct Persuasion* condition report the largest reductions of 23.8 minutes per day ($p < 0.001$). This suggests that while a DB intervention leads to less ambitious plans than with change talk focus, engaging with both the pros and cons before committing to change makes it easier to stick with the plan.

While all three treatments produce reductions in social media use two weeks later, consistent with the increase in motivation for time reduction observed in the main experiment—we do not see a monotonic relationship between the increase in motivation and the reduction in social media time use. Notably, while *Change Talk* and *Direct Persuasion* increase motivation to reduce social media screen time to a similar extent both immediately after the intervention and in the follow-up study more than two weeks later, the *Direct Persuasion* treatment had a more than twice as large effect on reported screen time, a substantial difference that is also statistically significant ($p < 0.05$). This pattern is consistent with the treatment effects on intended social media use and suggests that the intervention's stronger emphasis on concrete actions and strategies for translating intentions into actions—documented in the interview transcripts in Figure B.10—may be particularly effective at the final stage of behavior change, especially given the weaker effects of *Direct Persuasion* on some belief outcomes. Furthermore, while *Change Talk* led to a much higher effect in motivation than *Decisional Balance*, the observed treatment effect on time use is virtually identical across the two treatments ($p = 0.896$), despite respondents in the *Change Talk* treatment planning to reduce their social media consumption significantly more.

Alignment between actual and ideal time spent A natural question is whether the treatments help participants achieve their own stated preferences for social media use. To examine this, we construct binary indicators for whether participants' self-reported social media time in the follow-up survey falls within various thresholds of their pre-treatment ideal social media use. Table A.7 presents the results.

Direct Persuasion generates the largest and most consistent effects across all specifications: participants are 5.3 percentage points more likely to report actual use within 5 minutes of their ideal ($p < 0.01$), 7.0 percentage points more likely within 10 minutes ($p < 0.01$), 8.2 percentage points more likely within 20 minutes ($p < 0.01$), and 9.9 percentage points more likely to report use below their ideal plus 30 minutes ($p < 0.01$). Relative to control group means of 7.9–28.7%, these represent economically meaningful increases of 34–67%. *Change Talk* also increases alignment, though effects are smaller and less precisely estimated: 3.3 percentage points within 5 minutes ($p < 0.10$), 5.5 percentage points within 20 minutes ($p < 0.05$), and 5.1 percentage points below 30 minutes ($p < 0.10$). *Decisional Balance* produces no statistically significant effect on alignment at any threshold.

More broadly, if we interpret participants' pre-treatment ideal social media time as reflecting their *underlying* preferences for social media use, these findings suggest that our conversational interventions—particularly *Direct Persuasion*—increased participant welfare by reducing the gap between desired and actual behavior. This interpretation is consistent with models of self-control problems in which individuals persistently overconsume tempting goods relative to their own normative benchmarks (Allcott et al., 2022; O'Donoghue and Rabin, 1999).

Screenshot data To mitigate concerns about potential biases from self-reported social media time, we collected screenshot data on actual social media time for a subsample of respondents with Screen Time activated on their iPhone. Screen Time is an iPhone feature that records weekly data on the time spent on social media apps during the current week and the previous three weeks.[12] We offer respondents $2 for uploading screenshots of the weekly *Social* time for the previous two weeks.[13]

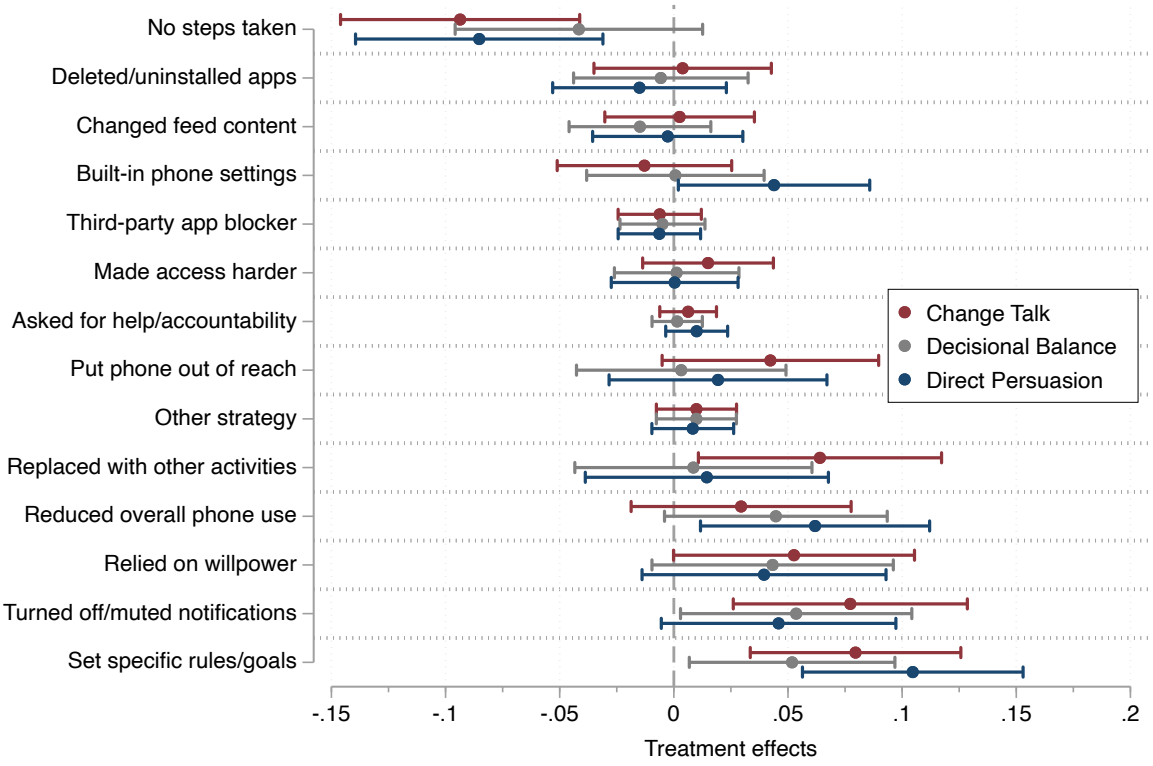
The *Social* category of the Screen Time app provides a verified measure of actual social media time, but it does not correspond perfectly to our self-reported measure of social media time. While the *Social* category includes four of the six social media apps used to elicit self-reported usage (Facebook, Instagram, Snapchat, TikTok, X/Twitter), it excludes two of the apps from our definition (YouTube and Reddit). Furthermore, it includes many apps not included in our definition of social media (or the definition used in e.g. Allcott et al., 2022), such as messaging apps like WhatsApp and Messenger. We therefore interpret this measure as a verified but noisy measure of the social media screen time we are interested in.

Although 1,194 participants had Screen Time enabled at baseline, only 444 participants uploaded valid screenshots from the previous two weeks. To examine the reliability of the self-reported data, we examine the correlation between the self-reported social media screen

[12]Screen Time is shown disaggregated by app categories. We explicitly ask for a screenshot of the screen time spent on apps that fall in the *Social* category in the weekly Screen Time summary.

[13]We ask for screenshots of week 52 of December 2025 and week 1 of December 2026. The data collection for the main experiment was completed in week 51 of 2025, giving us two full weeks of post-treatment social media screen time for respondents who correctly upload screenshots for these weeks. We also ask for screenshots of week 51, but we do not use this week for validation as it overlaps with the data collection period for the main experiment.

Figure 5: Adoption of actions and strategies for reducing social media time



Note: This figure reports treatment effect estimates on the use of different actions and strategies for reducing social media use using data from the follow-up survey. For each of the actions and strategies indicated in this figure, participants indicate whether they have used this strategy to reduce their social media use in the two weeks since the main survey. We regress a binary indicator having adopted a given strategy on treatment indicators and our preregistered set of control variables. 95% confidence intervals based on robust standard errors are shown.

time data and the screenshot data. This correlation—despite measurement differences across the two constructs—is substantial at 0.53 and highly statistically significant ($p < 0.001$). This suggests the self-reported data does not suffer from large reporting biases.

While we lack the statistical power to estimate treatment effects precisely for this limited subsample, we can test whether treatment effects on self-reported social media time differ significantly from treatment effects on iOS's Screen Time data through a seemingly unrelated regression (SUR) of both self-reported and verified social media screen time among the subsample who uploaded screenshots. We cannot reject equality of treatment effects across these two outcomes ($p > 0.100$), bolstering the reliability of the self-reported data. However, these results should be taken with some caution as treatment effect estimation becomes very noisy when we focus on the subsample that uploaded screenshots.

4.4.3 Heterogeneity

It is natural to hypothesize that the effectiveness of different conversational protocols at inducing behavioral change might plausibly depend on the perceived misalignment between people's desired level of social media use and the time they actually spent on social media. To examine this dimension of heterogeneity, we calculate the gap between participants' ideal and actual social media time as reported prior to our intervention. Table A.8 presents treatment effect estimates separately for participants with above-median ("high gap") and below-median ("low gap") misalignment between their ideal and actual social media time.

In Columns 1 and 2, we focus on predicted social media time first. Treatment effects are significantly larger among participants with a high gap between actual and ideal social media time, with effect sizes more than doubling. Among participants with a low gap, only *Change Talk* and *Direct Persuasion* induce changes in predicted future use. When turning to reported social media time in Columns 3 and 4, patterns are broadly similar. Here, only *Direct Persuasion* leads to reduction in reported social media time among participants with a low gap, while all treatments lead to sizable reductions among participants with a high baseline gap. This suggests that our treatments effectively capitalize on an existing perception of self-control issues, and use the conversational space to turn this perception into a source of behavior change.[14]

4.4.4 Strategies for change

Given the reductions in self-reported social media use documented above, a natural question is whether treatments differentially affected the actions and strategies participants adopted to achieve these reductions. In the follow-up survey, we included a multiple choice battery asking participants to select all strategies that they have used in the previous two weeks to reduce their social media time (e.g., "turned off notifications").[15]

Before turning to treatment effects, we provide descriptive evidence on how intentions to use a strategy translate into actual adoption of a strategy. Specifically, we ask whether participants are more likely to report having relied on strategies that they previously discussed with the chatbot during our main survey. This exercise also helps validate the quality of self-reported data on strategy use. We find that the probability that a participant reports having used a given strategy increases by 12.6 percentage points ($p < 0.001$) if this strategy was also discussed in the conversation with the AI interviewer during the main survey. This is a sizable increase of 122% compared to the 10.3% baseline probability that participants report using a given strategy if the strategy was *not* mentioned during the AI interview.

[14] Appendix Table A.9 shows that we obtain qualitatively similar but less pronounced heterogeneity in treatment effects when instead splitting the sample on pre-treatment time spent on social media.

[15] The battery includes 12 concrete strategies in addition to a "No steps taken" and "Other strategy" option. The concrete strategies coincide with the strategies we identified in a systematic review of our interview transcripts and were previously discussed in Section 3.3.2 above.

We next turn to treatment effects on the adoption of different reduction strategies. Figure 5 presents disaggregated treatment effects for each individual strategy. We document several patterns consistent with the differential behavioral effects documented above. First, both *Persuasion* and *Change Talk* reduce the likelihood of participants taking "no steps" at all to reduce their social media use by 8.5 and 9.4 percentage points ($p < 0.01$), respectively. The magnitude of this effect is sizable and corresponds to a 26% reduction in the probability of taking no actions in the control group. *Decisional Balance* produces a smaller reduction of 4.2 percentage points ($p > 0.100$). Second, we find that our conversational protocols increase the uptake of a broad range of strategies, consistent with the strong individual-level heterogeneity in how people approach the challenge of behavioral change. For example, *Direct Persuasion* generates the largest uptake of setting specific rules or goals for their social media use and using built-in phone settings—strategies closely linked to the directive conversational protocol that attempted to extract concrete commitments and measurable reduction targets from participants during the interviews. *Change Talk*, on the other hand, seems particularly effective in increasing the uptake strategies such as turning off notifications, replacing social media use with other activities, or setting specific rules or goals. Treatment effects on other strategies are more muted and generally similar across conditions.

To reduce noise in measurement, we aggregate strategies into two broader categories: *behavioral strategies* (e.g., set specific rules/goals, put phone out of reach, reduced overall phone use, replaced with other activities, asked for help/accountability, relied on willpower, made access harder) and *technology-based strategies* (e.g., Built-in phone settings, third-party app blocker, turned off/muted notifications, deleted/uninstalled apps, changed feed content). Table 2 reports treatment effects on a dummy for using at least one strategy from each category. We find that all three treatments increase the uptake of behavioral strategies (Column 1), with larger effects for *Change Talk* (+11.0 pp) and *Direct Persuasion* (+9.3 pp). The treatments differ, however, in their effects on the use of technology-based strategies. Participants in *Direct Persuasion* are 7.3 pp more likely to rely on technology to reduce their social media use ($p < 0.001$)—consistent with the disaggregated effects documented above. *Decisional Balance*, by contrast, does not increase the uptake of technological solutions ($p > 0.100$), and *Change Talk* only marginally increases the use of technological solutions. These results imply that *Direct Persuasion* may be more effective at decreasing actual social media use relative to the other treatments by promoting greater take-up of technological aids—which are less reliant on participants' moment-to-moment self-control.

4.4.5 Experimenter demand effects

To assess whether experimenter demand is likely to bias our treatment effects, we elicited participants' beliefs about the study's purpose via an open-ended question at the end of our main survey (Bursztyn et al., 2025, forthcoming). We classify answers into 5 categories: (i)

Table 2: Treatment effects on actual actions and strategies used to reduce social media screen time

	(1) No steps taken	(2) Behavioral strategy used	(3) Technolgy-based strategy used
Change Talk (a)	-0.094***	0.110***	0.054*
	(0.027)	(0.029)	(0.030)
Decisional Balance (b)	-0.042	0.072**	0.022
	(0.028)	(0.030)	(0.029)
Direct Persuasion (c)	-0.085***	0.093***	0.073**
	(0.028)	(0.030)	(0.030)
Observations	2,115	2,115	2,115
R^2	0.036	0.026	0.053
Control group mean	0.321	0.577	0.384
Controls	Yes	Yes	Yes
p-value: a=b	0.046	0.182	0.266
p-value: a=c	0.749	0.560	0.549
p-value: b=c	0.106	0.468	0.089

Note: This table presents treatment effect estimates using data from the follow-up survey. "Change Talk", "Decisional Balance" and "Direct Persuasion" are treatment indicators, with the time use interview control group being the omitted category. "No steps taken" takes value one if participants report to not have taken any actions or strategies to reduce their social media time between the main survey and the follow-up survey, and zero otherwise. "Behavioral strategy used" is a binary dummy for reporting to have used at least one of the following actions or strategies since the main survey: Set specific rules/goals, put phone out of reach, reduced overall phone use, replaced with other activities, asked for help/accountability, relied on willpower, made access harder. "Technology-based strategy used" is a binary dummy for reporting to have used at least one of the following actions or strategies since the main survey: Built-in phone settings, third-party app blocker, turned off/muted notifications, deleted/uninstalled apps, changed feed content. We report p-values for two-sided tests of equality of coefficients at the bottom of the table. Robust standard errors are shown in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

correctly identifying the study as examining an AI chatbot's effect on social media use reduction, (ii) identifying only the social media reduction component, (iii) identifying only the chatbot component, (iv) proposing an entirely different hypothesis, or (v) indicating they did not know.[16]

A particularly relevant concern is that participants who believe the researchers expect a reduction in social media use may adjust their self-reports to conform to this perceived expectation. To assess whether such experimenter demand effects are present in the data, we compare treatment effects across two subsamples: participants who believed that a reduction in social media use was expected and those who did not. Using seemingly unrelated regressions (SURs), we test whether treatment effects differ across these groups for our primary outcomes of motivation and perceived costs and benefits. We find no statistically significant differences in treatment effects for either outcome ($p = 0.29$ and $p = 0.18$, respectively), alleviating concerns about bias from experimenter demand. If anything, the pattern of point estimates runs counter to such concerns: participants who believed that a reduction in social media use was expected exhibit

[16]Figure B.14 shows the distribution of categories.

weaker treatment effects on motivation.

5 Concluding remarks

This paper employs LLMs to provide experimental evidence at scale on which types of conversations are most effective at motivating behavioral change. We study this question in the context of social media consumption, where most people have a strong desire to use social media less than they currently do.

Our results have several implications for behavioral change interventions. First, while reactance theory in psychology predicts that attempts to increase motivation through direct persuasion will be ineffective or even backfire (Steindl et al., 2015), we find that our *Direct Persuasion* treatment is most effective at reducing self-reported social media use, even though *Change Talk* is even more effective at increasing people's motivation for change. This effect is likely driven by the focus in the *Direct Persuasion* treatment on creating concrete and strict plans, suggesting that interventions that provide clear advice on successful behavioral change strategies are likely to be more effective for changing actual behavior than interventions that emphasize that people should design their own plan for change. Second, in line with current best-practice guidelines for Motivational Interviews (Miller and Rose, 2015), our findings demonstrate that focusing on the benefits of change is a more robust strategy for increasing motivation than approaches that emphasize decisional balance.

We also establish that behavioral interventions can be successfully delivered with LLMs, allowing researchers to conduct high-quality behavioral interventions at scale with high fidelity to the conversational protocols. More broadly, this paper provides a template for studying the effectiveness of different conversational protocols at scale and identifies the "active ingredients" that make them work, thus allowing researchers to open the previous black box of conversations as treatments. As AI capabilities continue to advance, such approaches may prove valuable not only for identifying the most effective behavioral interventions but also for delivering personalized behavioral interventions at scale.

References

Al-Ubaydli, Omar, John A. List, and Daniel Suskind, "What Can We Learn from Experiments? Understanding the Threats to the Scalability of Experimental Results," *American Economic Review*, 2017, *107* (5), 282–286.

Allcott, Hunt, "The welfare effects of misperceived product costs: Data and calibrations from the automobile market," *American Economic Journal: Economic Policy*, 2013, *5* (3), 30–66.

_ , Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow, "The welfare effects of social media," *American Economic Review*, 2020, *110* (3), 629–676.

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

Amrhein, Paul C., William R. Miller, Carolina E. Yahne, Michael Palmer, and Laura Fulcher, "Client commitment language during motivational interviewing predicts drug use outcomes," *Journal of Consulting and Clinical Psychology*, 2003, *71* (5), 862.

Apodaca, Timothy R. and Richard Longabaugh, "Mechanisms of change in Motivational Interviewing: A review and preliminary evaluation of the evidence," *Addiction*, 2009, *104* (5), 705–715.

Armantier, Olivier, Giorgio Topa, Wilbert van der Klaauw, and Basit Zafar, "An Overview of the Survey of Consumer Expectations," *Economic Policy Journal*, 2017.

Ash, Elliott and Stephen Hansen, "Text Algorithms in Economics," *Annual Review of Economics*, 2023, *15*, 659–688.

_ , Sergio Galletta, and Giacomo Opocher, "BallotBot: Can AI Strengthen Democracy?," *SSRN Electronic Journal*, 2025.

Auxier, Brooke and Monica Anderson, "Social Media Use in 2021," Pew Research Center April 2021. `https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/` (accessed August 22, 2023).

Bandura, Albert, "Self-efficacy: Toward a Unifying Theory of Behavioral Change," *Psychological Review*, 1977, *84* (2), 191–215.

Benartzi, Shlomo and Richard H. Thaler, "Heuristics and Biases in Retirement Savings Behavior," *Journal of Economic Perspectives*, 2007, *21* (3), 81–104.

Blattman, Christopher, Julian C. Jamison, and Margaret Sheridan, "Reducing crime and violence: Experimental evidence from cognitive behavioral therapy in Liberia," *American Economic Review*, 2017, *107* (4), 1165–1206.

Boissin, Elisa, Gordon Pennycook, David G. Rand et al., "Dialogues with Large Language Models Reduce Conspiracy Beliefs Even When the AI Is Perceived as Human," *PNAS Nexus*, 2025, *4* (1), pgaf325.

Bordalo, Pedro, Nicola Gennaioli, Giacomo Lanzani, and Andrei Shleifer, "A Cognitive Theory of Reasoning and Choice," Working Paper 33466, National Bureau of Economic Research 2025.

Braghieri, Luca, Ro'ee Levy, and Alexey Makarin, "Social media and mental health," *American Economic Review*, 2022, *112* (11), 3660–3693.

_ , Sarah Eichmeyer, Ro'ee Levy, Markus Mobius, Jacob Steinhardt, and Ruiqi Zhong, "Article-Level Slant and Polarization of News Consumption on Social Media," *Working Paper*, 2025.

Brehm, Jack W., *A Theory of Psychological Reactance*, Cambridge, MA: Academic Press, 1966. Foundational formulation of psychological reactance theory.

Burke, Brian L., Hal Arkowitz, and Manuel Menchola, "The Efficacy of Motivational Interviewing: A Meta-Analysis of Controlled Clinical Trials," *Journal of Consulting and Clinical Psychology*, 2003, *71* (5), 843–861.

Bursztyn, Leonardo, Benjamin Handel, Rafael Jiménez-Durán, and Christopher Roth, "When product markets become collective traps: The case of social media," *American Economic Review*, 2025, *115* (12), 4105–4136.

_ , Ingar Haaland, Nicolas Roever, and Christopher Roth, "The Social Desirability Atlas," *Journal of Political Economic Micro*, 2025, forthcoming.

Carey, K. B., M. P. Carey, S. A. Maisto, and J. M. Henson, "Brief motivational interventions for heavy college drinkers: A randomized controlled trial," *Journal of Consulting and Clinical Psychology*, 2006, *74* (5), 943–954.

Celebi, Can, Christine Exley, Sören Harrs, Hannu Kivimaki, Marta Serra-Garcia, and Jeffrey Yusof, "Mission Possible: The Collection of High-Quality Data," 12 2025. Unpublished manuscript.

Chopra, Felix and Ingar Haaland, "Conducting Qualitative Interviews with AI," *CESifo Working Paper No. 10666*, 2023.

_ , Ingar K. Haaland, Fabian Roeben, Christopher Roth, and Vanessa Sticher, "News Customization with AI," CESifo Working Paper 12121, CESifo 2025.

Conlon, John J., "Memory Rehearsal and Belief Biases," Technical Report, Working paper 2025.

Costello, Thomas H., Gordon Pennycook, and David G. Rand, "Durably Reducing Conspiracy Beliefs through Dialogues with Artificial Intelligence," *Science*, 2024, *385* (6713), eadq1814.

Deci, Edward L. and Richard M. Ryan, *Intrinsic Motivation and Self-Determination in Human Behavior*, New York: Plenum Press, 1985.

DellaVigna, Stefano and Elizabeth Linos, "RCTs to Scale: Comprehensive Evidence from Two Nudge Units," *Econometrica*, 2022, *90* (1), 81–116.

_ and Matthew Gentzkow, "Persuasion: empirical evidence," *Annual Review of Economics, Volume 2*, 2010.

_ and Ulrike Malmendier, "Paying Not to Go to the Gym," *American Economic Review*, 2006, *96* (3).

Foster, Dawn W, Clayton Neighbors, and Ankita Pai, "Decisional Balance: Alcohol Decisional Balance Intervention for Heavy Drinking Undergraduates," *Substance Use & Misuse*, 2015, *50* (13).

Frost, Helen, Pauline Campbell, Margaret Maxwell, Ronan E. O'Carroll, Stephan U. Dombrowski, Brian Williams, Helen Cheyne, Emma Coles, and Alex Pollock, "Effectiveness of Motivational Interviewing on adult behaviour change in health and social care settings: A systematic review of reviews," *PLOS ONE*, 2018, *13* (10), e0204890.

Galasso, Vincenzo, Tommaso Nannicini, and Debora Nozza, "We Need to Talk: Audio Surveys and Information Extraction," *Working Paper*, 2024.

Geiecke, Friedrich and Xavier Jaravel, "Conversations at scale: Robust ai-led interviews with a simple open-source platform," *Available at SSRN 4974382*, 2025.

Godin, Gaston and Mark Conner, "Intention-behavior relationship based on epidemiologic indices: an application to physical activity," *American Journal of Health Promotion*, 2008, *22* (3).

Graeber, Thomas, Christopher Roth, and Constantin Schesch, "Explanations," Technical Report, CESifo Working Paper 2024.

Grootendorst, Maarten, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv preprint arXiv:2203.05794*, 2022.

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

_ , _ , Stefanie Stantcheva, and Johannes Wohlfart, "Understanding Economic Behavior Using Open-Ended Survey Data," *Journal of Economic Literature*, 2025, *63* (4), 1244–1280.

Habibi, Mahyar, Dirk Hovy, and Carlo Schwarz, "The Content Moderator's Dilemma: Removal of Toxic Content and Distortions to Online Discourse," *arXiv preprint arXiv:2412.16114*, 2024.

Haidt, Jonathan, *The anxious generation: How the great rewiring of childhood is causing an epidemic of mental illness*, Penguin, 2024.

Hallgren, Kevin A., Barbara S. McCrady, Elizabeth E. Epstein, Scott Cook, Nicole K. Jensen, and Timothy Hildebrandt, "Variability in Motivational Interviewing Adherence Across Sessions, Providers, Sites, and Research Contexts," *Journal of Substance Abuse Treatment*, 2018, *84*, 20–26.

Janis, Irving L. and Leon Mann, *Decision Making: A Psychological Analysis of Conflict, Choice, and Commitment*, Free Press, 1977.

LaBrie, J. W., E. R. Pedersen, M. Earleywine, and H. Olsen, "Reducing heavy drinking in college males with the decisional balance: Analyzing an element of Motivational Interviewing," *Addictive Behaviors*, 2006, *31* (2), 254–263.

List, John A., *The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale*, New York: Currency, 2022.

_ , "Optimally generate policy-based evidence before scaling," *Nature*, 2024, *628*.

Ludwig, Jens, Sendhil Mullainathan, Sophia Pink, and Ashesh Rambachan, "Algorithms as a Vehicle to Reflective Equilibrium: Behavioral Economics 2.0," 2025. Draft chapter for *The Economics of Transformative AI*.

Madson, Michael B., Rebecca S. Mohn, Julie A. Schumacher, and Angela Landry, "Measuring Client Experiences of Motivational Interviewing During a Lifestyle Intervention," *Measurement and Evaluation in Counseling and Development*, 2015, *48* (2), 140–151.

_ , Richard S. Mohn, Allan Zuckoff, Julie A. Schumacher, Jane Kogan, Shari Hutchison, Emily Magee, and Bradley Stein, "Measuring client perceptions of motivational interviewing: Factor analysis of the Client Evaluation of Motivational Interviewing scale," *Journal of Substance Abuse Treatment*, 2013, *44* (3), 330–335.

Mair, Jacqueline Louise, Alicia Salamanca-Sanabria, Mareike Augsburger, Bea Franziska Frese, Stefanie Abend, Robert Jakob, Tobias Kowatsch, and Severin Haug, "Effective Behavior Change Techniques in Digital Health Interventions for the Prevention or Management of Noncommunicable Diseases: An Umbrella Review," *Annals of Behavioral Medicine*, October 2023, *57* (10).

Miller, W. R., R. G. Benefield, and J. S. Tonigan, "Enhancing motivation for change in problem drinking: A controlled comparison of two therapist styles," *Journal of Consulting and Clinical Psychology*, 1993, *61* (3), 455–461.

Miller, William R. and Gary S. Rose, "Motivational Interviewing and Decisional Balance: Contrasting Responses to Client Ambivalence," *Behavioural and Cognitive Psychotherapy*, 2015, *43* (2), 129–141.

_ **and Stephen Rollnick**, *Motivational interviewing: Preparing people to change addictive behavior*, The Guilford Press, 1991.

Miller, William R and Stephen Rollnick, *Motivational Interviewing: Helping People Change and Grow*, Guilford press, 2012.

Miller, William R. and Stephen Rollnick, *Motivational Interviewing: Helping People Change*, 3rd ed., New York, NY: Guilford Press, 2013. MI conceptualizes ambivalence about change as a core target of the counseling approach.

Moyers, T.B., J.K. Manuel, and D. Ernst, "Motivational Interviewing Treatment Integrity Coding Manual 4.2.1," Technical Report, University of New Mexico, Center on Alcoholism, Substance Abuse, and Addictions (CASAA), Albuquerque, NM 2014. Unpublished manual. Revised June 2015.

Moyers, Theresa B., Lauren N. Rowell, Jennifer K. Manuel, Denise Ernst, and Jon M Houck, "The motivational interviewing treatment integrity code (MITI 4): rationale, preliminary reliability and validity," *Journal of Substance Abuse Treatment*, 2016, *65*, 36–42.

_ **, Tim Martin, Jon M. Houck, Paulette J. Christopher, and J. Scott Tonigan**, "From in-session behaviors to drinking outcomes: a causal chain for motivational interviewing," *Journal of Consulting and Clinical Psychology*, 2009, *77* (6), 1113–1124.

Mullainathan, Sendhil, "Economics in the Age of Algorithms," *AEA Papers and Proceedings*, May 2025, *115*, 1–23.

_ **and Ashesh Rambachan**, "Science in the Age of Algorithms," 2025. Draft chapter for *The Economics of Transformative AI*.

O'Donoghue, Ted and Matthew Rabin, "Doing It Now or Later," *American Economic Review*, 1999, *89* (1), 103–124.

Prochaska, James O. and Carlo C. DiClemente, "Stages and Processes of Self-Change of Smoking: Toward an Integrative Model of Change," *Journal of Consulting and Clinical Psychology*, 1983, *51* (3), 390–395.

Rilla, Raluca, Tobias Werner, Hiromu Yakura, Iyad Rahwan, and Anne-Marie Nussberger, "Recognising, Anticipating, and Mitigating LLM Pollution of Online Behavioural Research," 2025.

Samdal, Gro Beate, Geir Egil Eide, Tom Barth, Geoffrey Williams, and Eivind Meland, "Effective behaviour change techniques for physical activity and healthy eating in overweight and obese adults; systematic review and meta-regression analyses," *Int J Behav Nutr Phys Act*, 2017, *14* (1), 42.

Steindl, Christina, Eva Jonas, Sandra Sittenthaler, Eva Traut-Mattausch, and Jeff Greenberg, "Understanding Psychological Reactance: New Developments and Findings," *Zeitschrift für Psychologie*, 2015, *223* (4).

Watson, David, Lee Anna Clark, and Auke Tellegen, "Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales," *Journal of Personality and Social Psychology*, 1988, *54* (6), 1063–1070.

Zhu, SuFen, Deepra Sinha, Megan Kirk, Moscho Michalopoulou, Anisa Hajizadeh, Gina Wren, Paul Doody, Lucy Mackillop, Ralph Smith, Susan A Jebb et al., "Effectiveness of behavioural interventions with motivational interviewing on physical activity outcomes in adults: systematic review and meta-analysis," *bmj*, 2024, *386.*

For online publication only:

Evaluating Behavioral Interventions at Scale with AI

Felix Chopra Ingar Haaland Nicolas Roever Christopher Roth

January 17, 2026

Summary of the Online Appendix

The Online Appendix provides supplementary descriptive evidence, robustness analyses, and implementation details.

- Appendix A contains additional tables.
- Appendix B presents additional figures.
- Appendix C provides additional analyses.
- Appendix D provides example interviews.
- Appendix E contains the experimental instructions.

A Additional Tables

Table A.1: Summary statistics and test of balance

	(1) Sample mean	(2)	(3)	(4)	(5)	(6)	(7)
		Mean differences and p-values					
		C vs T1	C vs T2	C vs T3	T1 vs T2	T1 vs T3	T2 vs T3
Age	41.034	0.339	0.515	0.810	0.176	0.471	0.295
	(12.860)	(0.620)	(0.458)	(0.256)	(0.797)	(0.502)	(0.679)
Female	0.599	0.052*	0.020	-0.001	-0.031	-0.053**	-0.021
	(0.490)	(0.052)	(0.449)	(0.972)	(0.233)	(0.047)	(0.427)
College degree	0.620	0.031	0.039	0.025	0.008	-0.005	-0.014
	(0.485)	(0.243)	(0.139)	(0.340)	(0.751)	(0.834)	(0.599)
Employed	0.672	0.032	0.012	-0.031	-0.020	-0.063**	-0.043*
	(0.469)	(0.206)	(0.646)	(0.225)	(0.421)	(0.013)	(0.093)
Log household income	11.050	0.049	0.021	-0.000	-0.028	-0.050	-0.022
	(0.768)	(0.240)	(0.613)	(0.995)	(0.490)	(0.229)	(0.602)
Caucasian/White	0.741	0.033	-0.020	0.002	-0.053**	-0.031	0.022
	(0.438)	(0.158)	(0.408)	(0.946)	(0.025)	(0.178)	(0.369)
African American/Black	0.125	-0.025	-0.002	-0.013	0.024	0.012	-0.012
	(0.331)	(0.155)	(0.923)	(0.464)	(0.184)	(0.491)	(0.523)
Hispanic origin	0.058	0.003	0.004	-0.011	0.001	-0.014	-0.015
	(0.235)	(0.830)	(0.782)	(0.371)	(0.950)	(0.265)	(0.241)
Country: United States	0.473	-0.002	-0.014	0.000	-0.011	0.003	0.014
	(0.499)	(0.932)	(0.615)	(0.992)	(0.674)	(0.924)	(0.607)
Social media use (min/day)	183.016	-5.103	-7.822	-6.847	-2.719	-1.744	0.975
	(110.931)	(0.401)	(0.196)	(0.267)	(0.644)	(0.771)	(0.870)
iPhone user	0.535	0.006	0.022	-0.015	0.016	-0.022	-0.037
	(0.499)	(0.813)	(0.410)	(0.579)	(0.554)	(0.426)	(0.166)
TikTok or Instagram user	0.918	-0.026*	-0.003	-0.017	0.022	0.009	-0.013
	(0.274)	(0.086)	(0.804)	(0.250)	(0.139)	(0.568)	(0.365)
p-value of joint F-test		0.243	0.827	0.852	0.355	0.121	0.706
Observations	2,719	1,360	1,351	1,350	1,369	1,368	1,359

Note: This balance table presents the mean differences and p-values (in brackets) for a range of background variables across different treatment groups in column 2-7. Column 1 shows the full sample mean of baseline covariates and the associated standard deviation. Column 2 presents mean differences between the control group and *Change Talk*. Columns 3 and 4 present analogous differences between the control group and *Decisional Balance* and *Direct Persuasion*. Columns 5-7 present differences between the respective treatment arms indicated by the column header. "Age" is the respondents' numerical age. "Female" is a binary indicator taking value one for female respondents. "College degree" is a binary indicator for having completed a college degree. "Employed" is a binary indicator for being employed. "Log household income" is the log of the midpoint of the participant's household income. "Caucasian/White" and "African American/Black" are indicators for the respective groups. "Hispanic origin" is a binary indicator for respondents of Hispanic origin. "Country: United States" is a dummy for residing in the US. "iPhone user" is a dummy for the main mobile device being an iPhone. "Social media time (min/day)" is the average number of minutes per day that the participant uses social media apps. The p-values of the joint F-test are determined by regressing the treatment indicator on the vector of covariates. The F-test tests the joint hypothesis that none of the covariates predict treatment assignment.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.2: Analysis of attrition for the follow-up survey

	Completed follow-up
	(1)
Change Talk	0.012
	(0.024)
Decisional Balance	0.002
	(0.024)
Direct Persuasion	-0.041
	(0.025)
Observations	2,719
R^2	0.002
Control group mean	0.712

Note: This table presents an analysis of attrition. We regress a dummy for completing the follow-up survey on treatment indicators. Robust standard errors are shown in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.3: Summary statistics and test of balance for the follow-up survey

	(1) Sample mean	(2)	(3)	(4)	(5)	(6)	(7)
		Mean differences and p-values					
		C vs T1	C vs T2	C vs T3	T1 vs T2	T1 vs T3	T2 vs T3
Age	42.130	0.089	-0.166	1.054	-0.255	0.965	1.220
	(12.929)	(0.907)	(0.833)	(0.197)	(0.741)	(0.227)	(0.138)
Female	0.602	0.039	0.012	0.004	-0.027	-0.035	-0.007
	(0.490)	(0.192)	(0.699)	(0.891)	(0.360)	(0.250)	(0.806)
College degree	0.617	0.024	0.030	0.022	0.006	-0.002	-0.008
	(0.486)	(0.420)	(0.310)	(0.461)	(0.827)	(0.959)	(0.791)
Employed	0.668	0.028	-0.000	-0.035	-0.028	-0.063**	-0.035
	(0.471)	(0.319)	(1.000)	(0.232)	(0.318)	(0.028)	(0.232)
Log household income	11.046	0.068	0.029	0.027	-0.040	-0.042	-0.002
	(0.765)	(0.147)	(0.542)	(0.581)	(0.380)	(0.375)	(0.969)
Caucasian/White	0.759	0.055**	-0.027	-0.004	-0.082***	-0.059**	0.023
	(0.428)	(0.028)	(0.322)	(0.886)	(0.001)	(0.020)	(0.404)
African American/Black	0.113	-0.036*	0.012	0.002	0.048**	0.037**	-0.011
	(0.316)	(0.051)	(0.540)	(0.934)	(0.010)	(0.044)	(0.601)
Hispanic origin	0.054	0.011	0.015	0.002	0.004	-0.008	-0.012
	(0.226)	(0.431)	(0.290)	(0.867)	(0.779)	(0.542)	(0.380)
Country: United States	0.447	0.001	-0.016	-0.000	-0.017	-0.001	0.015
	(0.497)	(0.970)	(0.608)	(0.991)	(0.579)	(0.962)	(0.620)
Social media use (min/day)	175.363	3.265	-1.741	-4.920	-5.006	-8.185	-3.179
	(108.354)	(0.626)	(0.793)	(0.463)	(0.448)	(0.220)	(0.630)
iPhone user	0.522	-0.007	0.046	-0.019	0.053*	-0.012	-0.065**
	(0.500)	(0.817)	(0.135)	(0.543)	(0.081)	(0.699)	(0.037)
TikTok or Instagram user	0.912	-0.024	-0.001	-0.027	0.022	-0.003	-0.026
	(0.284)	(0.163)	(0.929)	(0.123)	(0.191)	(0.860)	(0.145)
p-value of joint F-test		0.153	0.922	0.872	0.021	0.103	0.605
Observations	2,130	1,087	1,067	1,040	1,090	1,063	1,043

Note: This table presents a test of balance for the follow-up survey. This balance table presents the mean differences and p-values (in brackets) for a range of background variables across different treatment groups in column 2-7. Column 1 shows the full sample mean of baseline covariates and the associated standard deviation. Column 2 presents mean differences between the control group and *Change Talk*. Columns 3 and 4 present analogous differences between the control group and *Decisional Balance* and *Direct Persuasion*. Columns 5-7 present differences between the respective treatment arms indicated by the column header. "Age" is the respondents' numerical age. "Female" is a binary indicator taking value one for female respondents. "College degree" is a binary indicator for having completed a college degree. "Employed" is a binary indicator for being employed. "Log household income" is the log of the midpoint of the participant's household income. "Caucasian/White" and "African American/Black" are indicators for the respective groups. "Hispanic origin" is a binary indicator for respondents of Hispanic origin. "Country: United States" is a dummy for residing in the US. "iPhone user" is a dummy for the main mobile device being an iPhone. "Social media time (min/day)" is the average number of minutes per day that the participant uses social media apps. The p-values of the joint F-test are determined by regressing the treatment indicator on the vector of covariates. The F-test tests the joint hypothesis that none of the covariates predict treatment assignment.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.4: Treatment effects on motivation, beliefs and willingness to pay

	(1) Motivation (std.)	(2) Perceived cost of social media (std.)	(3) Social media makes life worse (std.)	(4) Awareness self-control problems (std.)	(5) Self-efficacy belief (std.)	(6) WTP ($)
Change Talk (a)	0.516***	0.451***	0.270***	0.216***	0.184***	0.816***
	(0.051)	(0.049)	(0.053)	(0.050)	(0.050)	(0.255)
Decisional Balance (b)	0.212***	0.127**	-0.155***	-0.013	0.140***	0.148
	(0.054)	(0.052)	(0.054)	(0.051)	(0.051)	(0.232)
Direct Persuasion (c)	0.432***	0.202***	0.021	0.078	0.191***	0.345
	(0.052)	(0.052)	(0.053)	(0.051)	(0.049)	(0.236)
Observations	2,719	2,719	2,719	2,719	2,719	2,719
R^2	0.092	0.101	0.073	0.106	0.068	0.037
Control group mean	0.000	0.000	-0.000	-0.000	-0.000	3.004
Controls	Yes	Yes	Yes	Yes	Yes	Yes
p-value: a=b	0.000	0.000	0.000	0.000	0.364	0.009
p-value: a=c	0.101	0.000	0.000	0.006	0.876	0.072
p-value: b=c	0.000	0.158	0.001	0.078	0.285	0.410

Note: This table presents treatment effect estimates using data from the main survey. "Change Talk", "Decisional Balance" and "Direct Persuasion" are treatment indicators, with the time use interview control group being the omitted category. The dependent variables in columns 1–5 are standardized to have a mean of zero and a standard deviation of one in the control group. The dependent variable in column 6 is measured in dollars. WTP is elicited using a multiple price list procedure. We use the midpoint of the resulting WTP interval as the dependent variable. For participants that always prefer the bonus or the app, we use $0 and $22.50 as their WTP. We report p-values for two-sided tests of equality of coefficients at the bottom of the table. Robust standard errors are shown in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.5: Treatment effects on ideal, predicted and actual social media time

	(1) Ideal social media time (min)	(2) Predicted social media time (min)	(3) Actual social media time (min)
Change Talk (a)	-7.754***	-25.265***	-11.203*
	(2.773)	(3.909)	(5.930)
Decisional Balance (b)	4.130	-14.552***	-11.937**
	(2.866)	(3.716)	(5.698)
Direct Persuasion (c)	-7.444**	-39.320***	-23.824***
	(2.959)	(4.213)	(6.181)
Observations	2,719	2,719	2,119
R^2	0.362	0.409	0.427
Control group mean	89.970	154.511	187.874
Controls	Yes	Yes	Yes
p-value: a=b	0.000	0.004	0.896
p-value: a=c	0.913	0.001	0.039
p-value: b=c	0.000	0.000	0.043

Note: This table presents treatment effect estimates using data from the main survey. Column 3 uses data from the follow-up survey. "Change Talk", "Decisional Balance" and "Direct Persuasion" are treatment indicators, with the time use interview control group being the omitted category. The dependent variables are measured in minutes per day and winsorized at the 5^{th} and 95^{th} percentile. We report p-values for two-sided tests of equality of coefficients at the bottom of the table. Robust standard errors are shown in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.6: Treatment effects on motivation and beliefs: Follow-up survey

	(1) Motivation (std.)	(2) Perceived cost of social media (std.)	(3) Social media makes life worse (std.)
Change Talk (a)	0.147**	0.167***	0.088
	(0.057)	(0.055)	(0.058)
Decisional Balance (b)	0.068	0.054	-0.073
	(0.059)	(0.057)	(0.059)
Direct Persuasion (c)	0.161***	0.118**	0.010
	(0.059)	(0.057)	(0.059)
Observations	2,130	2,126	2,126
R^2	0.067	0.084	0.084
Control group mean	-0.000	0.000	0.000
Controls	Yes	Yes	Yes
p-value: a=b	0.167	0.036	0.007
p-value: a=c	0.806	0.363	0.181
p-value: b=c	0.116	0.252	0.163

Note: This table presents treatment effect estimates using data from the follow-up survey. "Change Talk", "Decisional Balance" and "Direct Persuasion" are treatment indicators, with the time use interview control group being the omitted category. The dependent variables are standardized to have a mean of zero and a standard deviation of one in the control group. We report p-values for two-sided tests of equality of coefficients at the bottom of the table. Robust standard errors are shown in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.7: Treatment effects on the chance of implementing one's ideal social media time

	Actual time is close to ideal social media time (0/1)			
	(1) Within 5 min	(2) Within 10 min	(3) Within 20 min	(4) Within 30 min
Change Talk (a)	0.033*	0.029	0.055**	0.051*
	(0.018)	(0.020)	(0.024)	(0.026)
Decisional Balance (b)	0.011	0.009	0.020	0.040
	(0.017)	(0.019)	(0.024)	(0.027)
Direct Persuasion (c)	0.053***	0.070***	0.082***	0.099***
	(0.019)	(0.022)	(0.026)	(0.028)
Observations	2,119	2,119	2,119	2,119
R^2	0.036	0.045	0.112	0.123
Control group mean	0.079	0.109	0.202	0.287
Controls	Yes	Yes	Yes	Yes
p-value: a=b	0.222	0.318	0.159	0.683
p-value: a=c	0.319	0.065	0.312	0.096
p-value: b=c	0.031	0.006	0.019	0.041

Note: This table presents treatment effect estimates using data from the main and follow-up survey. The dependent variables are binary indicators for reporting an actual social media time that is within X minutes of participants' ideal social media time, as reported prior to our intervention. "Change Talk", "Decisional Balance" and "Direct Persuasion" are treatment indicators, with the time use interview control group being the omitted category. We report p-values for two-sided tests of equality of coefficients at the bottom of the table. Robust standard errors are shown in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.8: Heterogeneity in treatment effects by gap between actual and ideal social media time

	Predicted social time		Actual social time	
	(1) Low gap	(2) High gap	(3) Low gap	(4) High gap
Change Talk (a)	-12.513***	-31.942***	-2.728	-20.259**
	(3.993)	(5.050)	(7.656)	(9.054)
Decisional Balance (b)	-4.904	-19.114***	-4.594	-22.136**
	(3.864)	(5.187)	(7.271)	(8.899)
Direct Persuasion (c)	-21.046***	-52.803***	-19.723***	-30.256***
	(4.510)	(5.856)	(7.537)	(10.231)
Observations	1,361	1,358	1,105	1,014
R^2	0.630	0.457	0.413	0.309
Control group mean	114.712	191.451	139.191	239.198
Controls	Yes	Yes	Yes	Yes
p-value: a=b	0.036	0.011	0.806	0.818
p-value: a=c	0.000	0.000	0.040	0.387
p-value: b=c	0.000	0.000	0.040	0.387

Note: This table presents treatment effect estimates using data from the main and the follow-up survey. "Change Talk", "Decisional Balance" and "Direct Persuasion" are treatment indicators, with the time use interview control group being the omitted category. The dependent variables are measured in minutes per day. Columns with "Low gap" focus on the subsample of respondents with below-median gap between actual and ideal social media time, while columns with "High gap" focus on the subset with above-median gap. The gap is constructed by taking the difference between actual and ideal social media time reported pre-treatment. We report p-values for two-sided tests of equality of coefficients at the bottom of the table. Robust standard errors are shown in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.9: Heterogeneity in treatment effects by baseline social media screen time

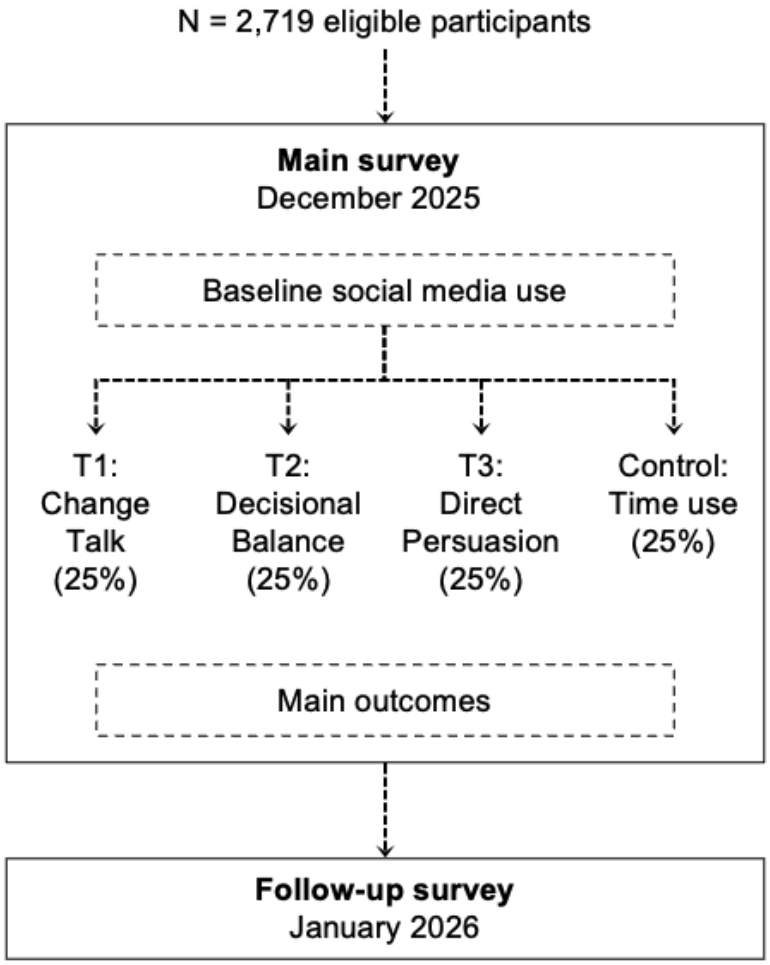
	Predicted social time		Actual social time	
	(1) Low use	(2) High use	(3) Low use	(4) High use
Change Talk (a)	-13.961***	-28.246***	-3.473	-16.542**
	(3.985)	(4.703)	(2.877)	(8.194)
Decisional Balance (b)	-6.340	-15.643***	-2.185	-20.608**
	(3.998)	(4.635)	(2.857)	(8.140)
Direct Persuasion (c)	-18.791***	-47.287***	-5.887**	-15.722*
	(4.588)	(5.294)	(2.852)	(8.906)
Observations	1,076	1,643	1,076	1,043
R^2	0.343	0.496	0.187	0.227
Control group mean	89.766	194.451	82.090	286.708
Controls	Yes	Yes	Yes	Yes
p-value: a=b	0.045	0.005	0.636	0.619
p-value: a=c	0.006	0.000	0.165	0.586
p-value: b=c	0.006	0.000	0.165	0.586

Note: This table presents treatment effect estimates using data from the main and the follow-up survey. "Change Talk", "Decisional Balance" and "Direct Persuasion" are treatment indicators, with the time use interview control group being the omitted category. The dependent variables are measured in minutes per day. Columns with "Low use" focus on the subsample of respondents with below-median baseline social media screen time, while columns with "High use" focus on the subset with above-median baseline social media use. We report p-values for two-sided tests of equality of coefficients at the bottom of the table. Robust standard errors are shown in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

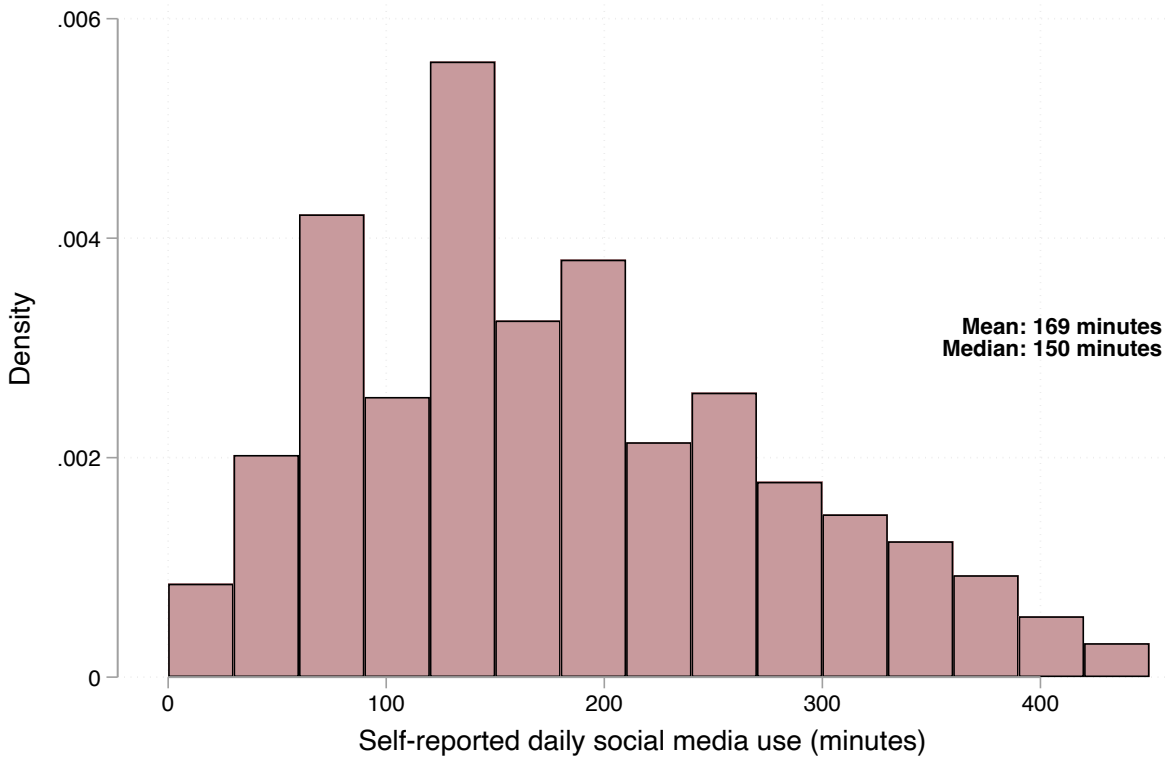
B Additional Figures

Figure B.1: Overview of the experimental design



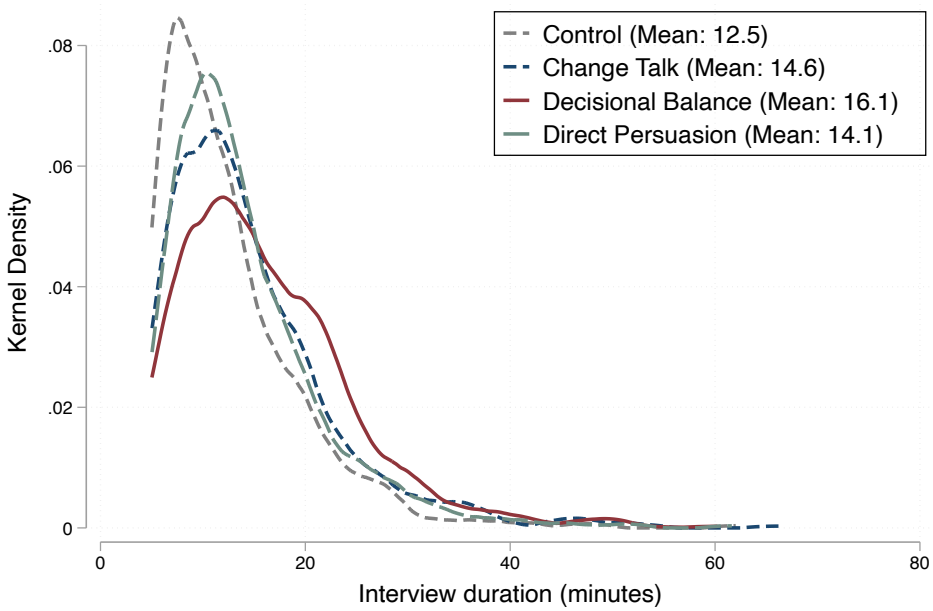
Note: This figure provides an overview of the experimental design. See Section 2 for details and Appendix Section E for the full experimental instructions.

Figure B.2: Distribution of self-reported social media use at baseline



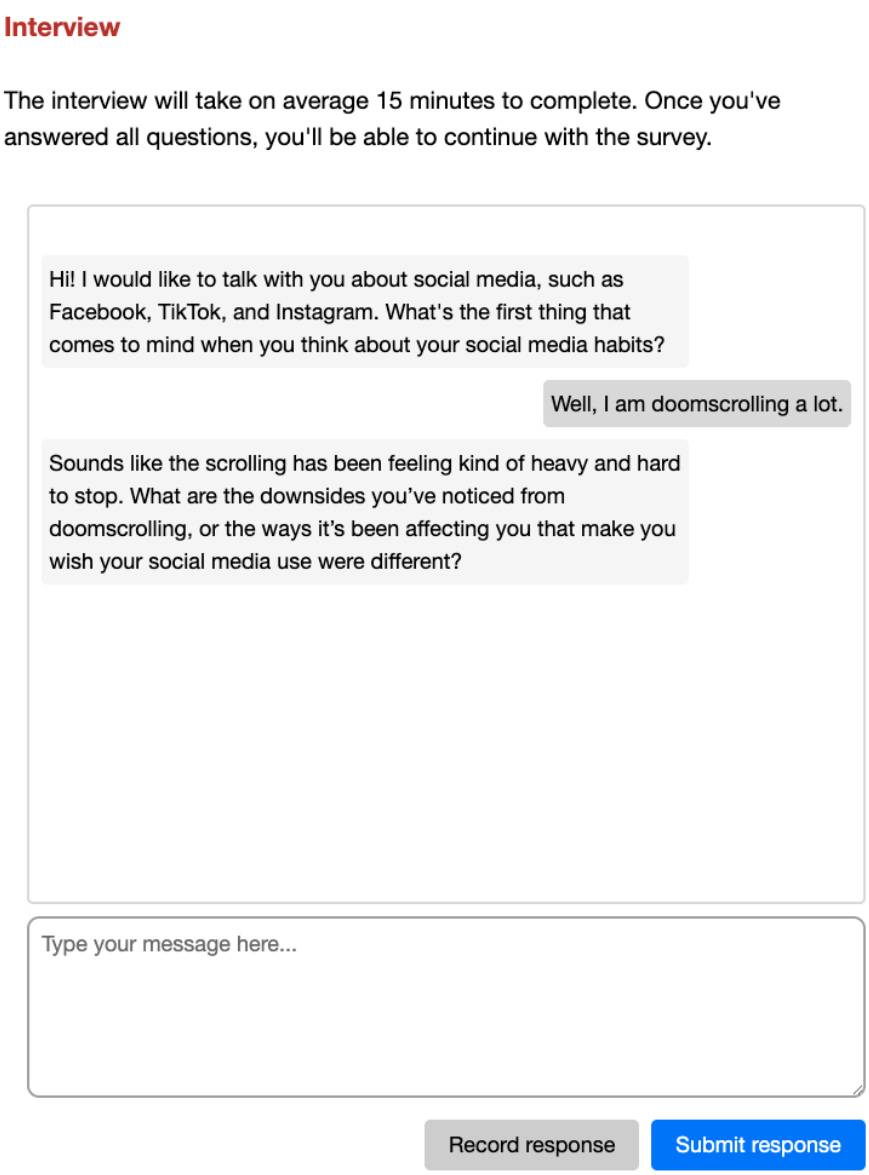
Note: This figure shows a histogram of the baseline distribution of reported social media use in minutes per day. For ease of presentation, we truncate the distribution at the 95th percentile.

Figure B.3: Interview duration by treatment group



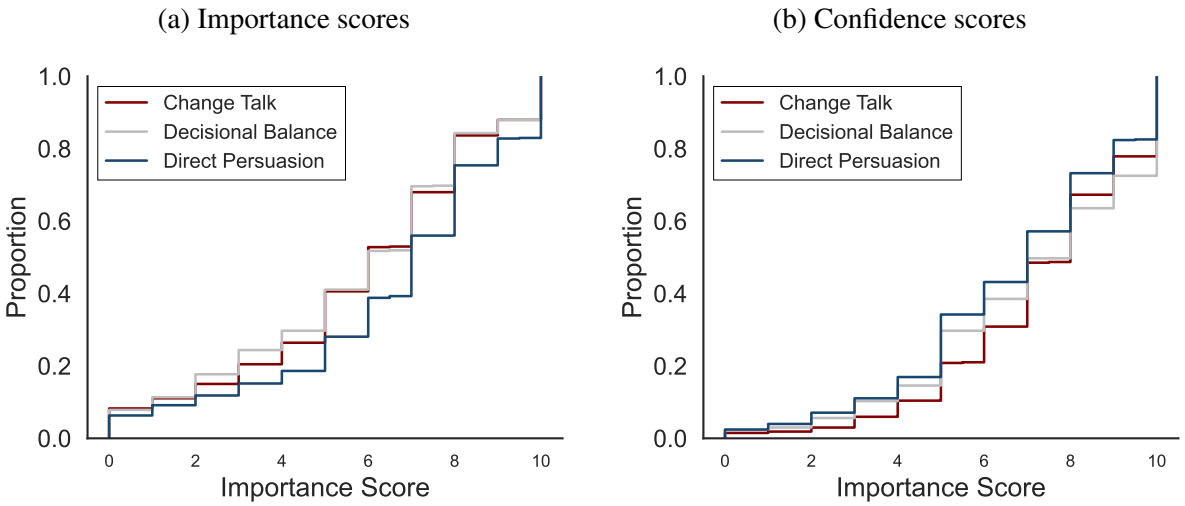
Note: This figure plots kernel density estimates of interview duration (in minutes) by treatment arm. Interview duration is measured as the total time participants spent interacting with the AI interviewer. Densities are estimated separately for each group using the same bandwidth.

Figure B.4: Chat interface embedded in our main survey



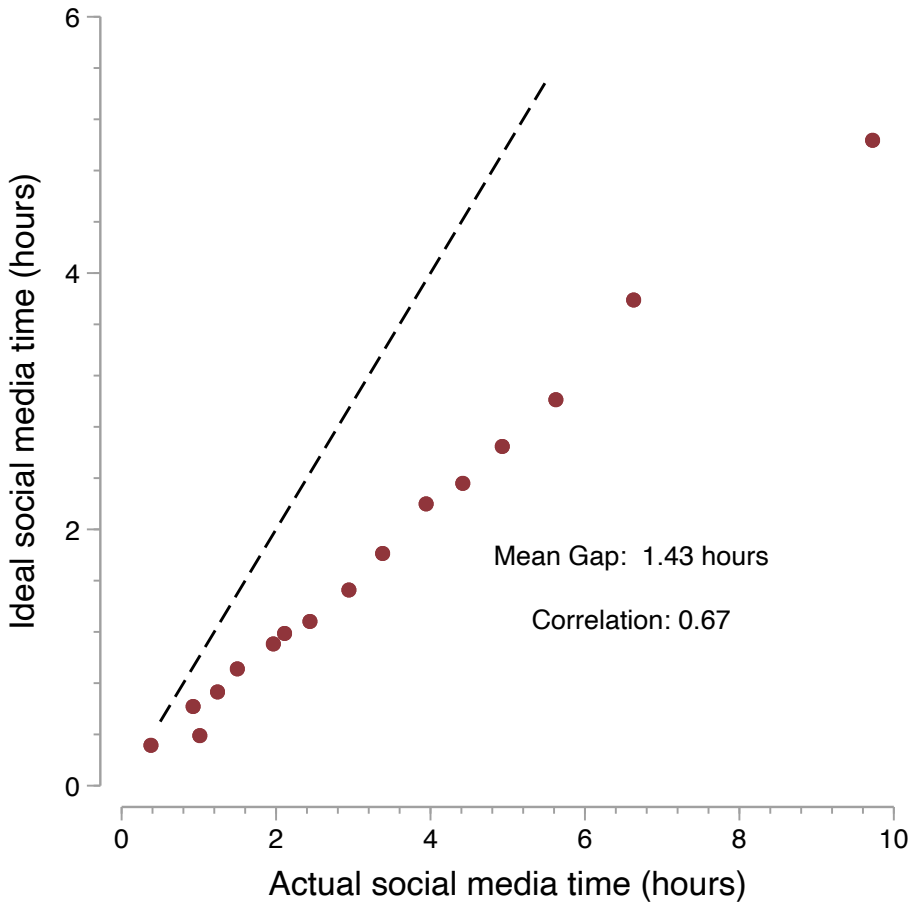
Note: This figure presents a screenshot of the chat interface embedded in our main survey for conducting interviews with an AI chatbot, which we obtain from Chopra and Haaland (2023). Participants can answer questions by typing or by recording a voice message that is automatically transcribed.

Figure B.5: Distribution of confidence and importance scores



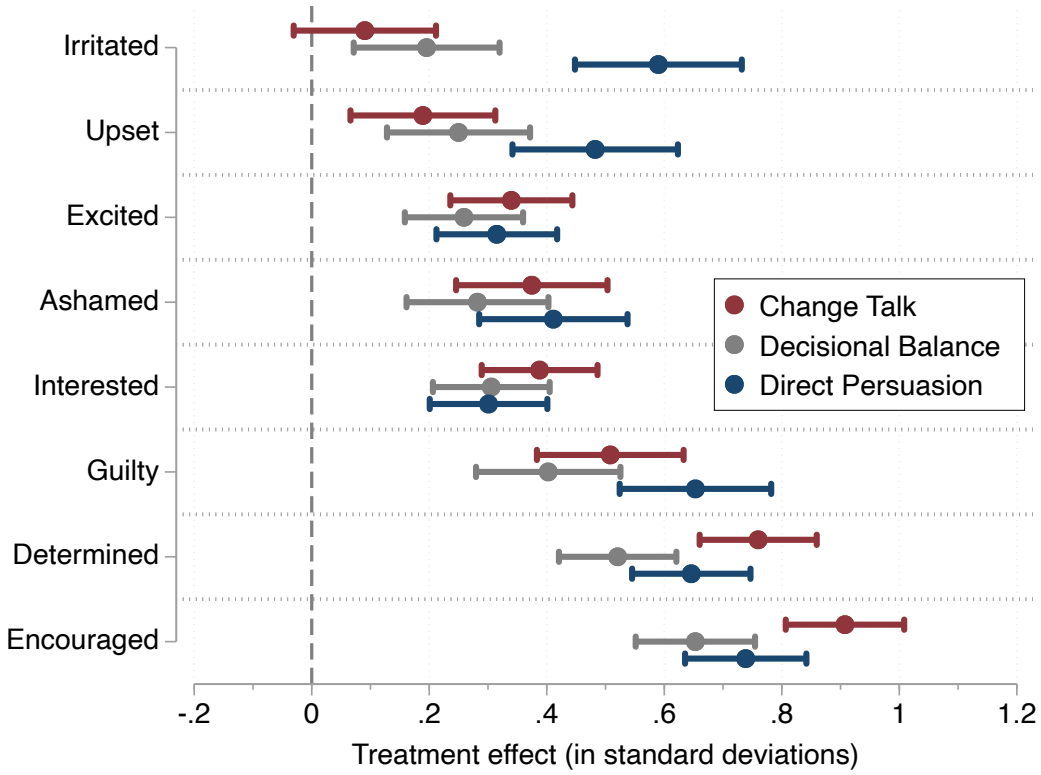
Note: Panel A plots the cumulative density function (CDF) of the importance scores reported in response to the scaling question on participants' perceived importance to reduce their social media use that were part of the AI-led interviews of our main survey. Panel B plots the CDF of the corresponding scores reported in response to the confidence scaling question. We present CDFs separately by treatment arm. Note that the scaling questions were not part of the control group interviews. Importance is measured by respondents' self-assessed importance of reducing social media use and confidence is measured by respondents' self-assessed confidence in their ability to reduce social media use, both elicited on an 11-point scale from 0 (lowest) to 10 (highest).

Figure B.6: Baseline social media use vs. baseline ideal social media use



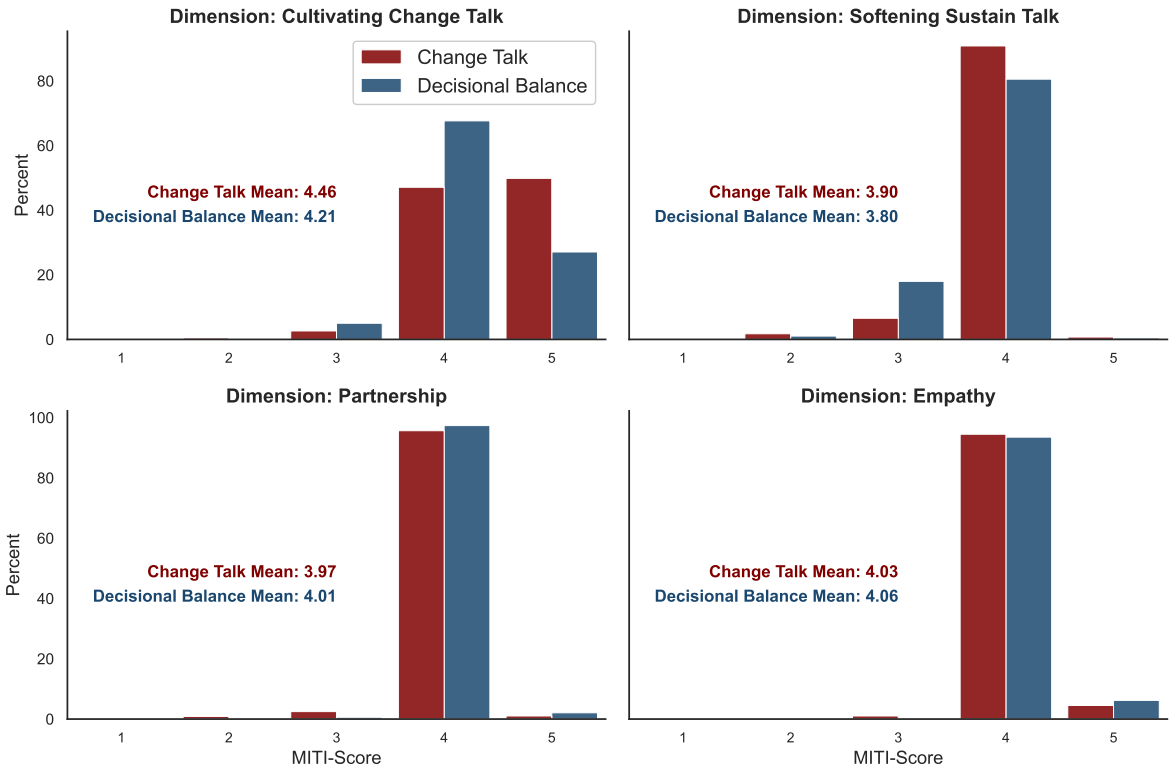
Note: This figure plots binned averages of participants' stated ideal daily social media time against their reported actual daily social media time prior to the intervention, both measured in hours per day. Each dot represents the mean ideal time within a bin of actual usage. The dashed 45-degree line indicates equality between actual and ideal time. The average gap between actual and ideal use is 1.43 hours, and the correlation between actual and ideal time is 0.671.

Figure B.7: Treatment effects on emotional states during the interviews



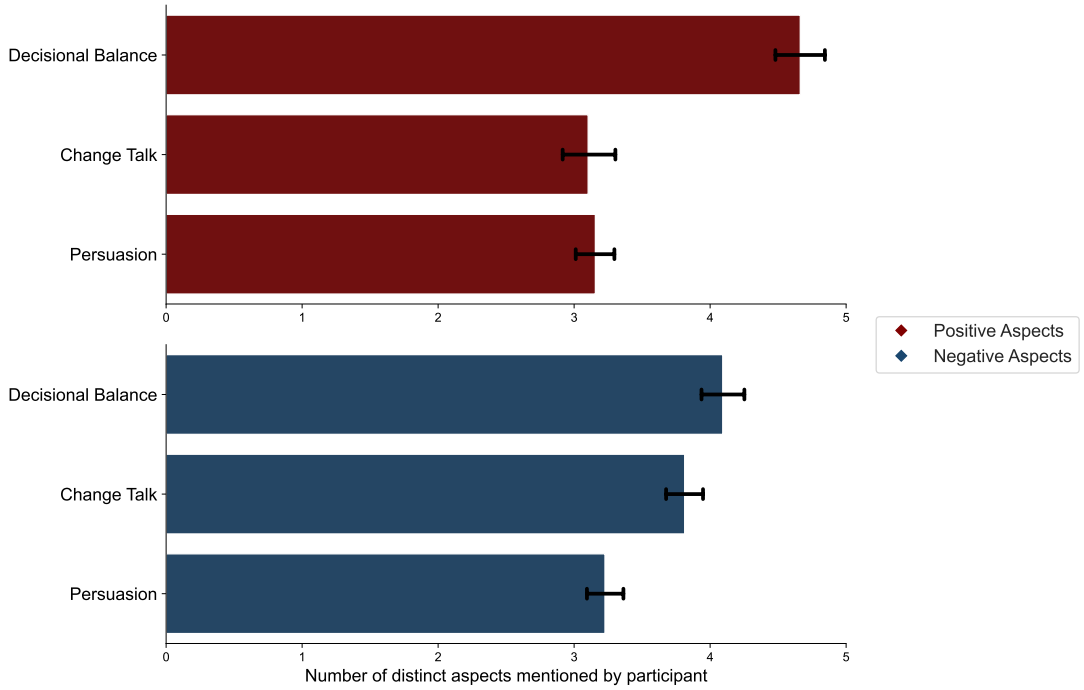
Note: This figure reports treatment effect estimates using data from the main survey. "Change Talk", "Decisional Balance" and "Direct Persuasion" are treatment indicators, with the time use interview control group being the omitted category. The dependent variables in each regression are standardized to have a mean of zero and a standard deviation of one among control group respondents. The dependent variables are self-reported emotional states experienced during the conversations: interested, excited, upset, guilty, irritated, ashamed, determined, and encouraged. Emotions are elicited on a 5-point Likert scale (very slightly or not at all; a little; moderately; quite a bit; extremely). 95% confidence intervals derived from robust standard errors are shown.

Figure B.8: MITI scores for MI interviews by dimension



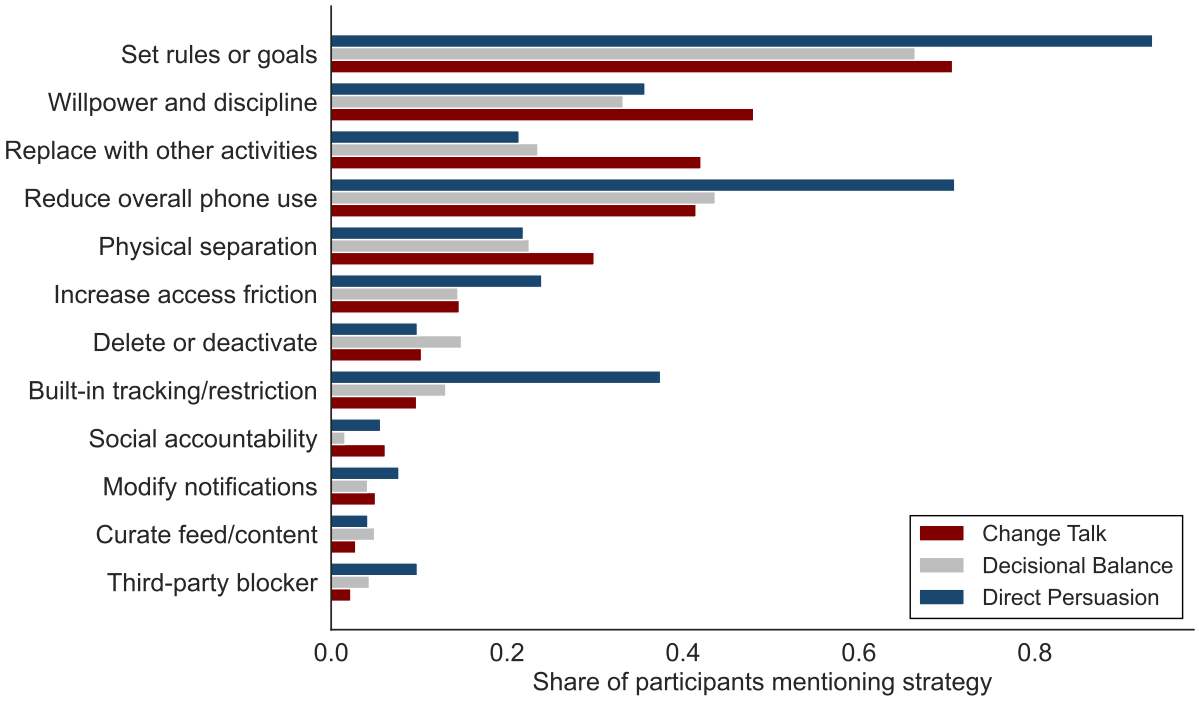
Note: This figure presents histograms for the four dimensions of the global MITI-score for all interviews in the treatment conditions *Change Talk* and *Decisional Balance*. We obtain the score using the workflow described in Section C.1.

Figure B.9: Positive and negative aspects of social media mentioned by participants



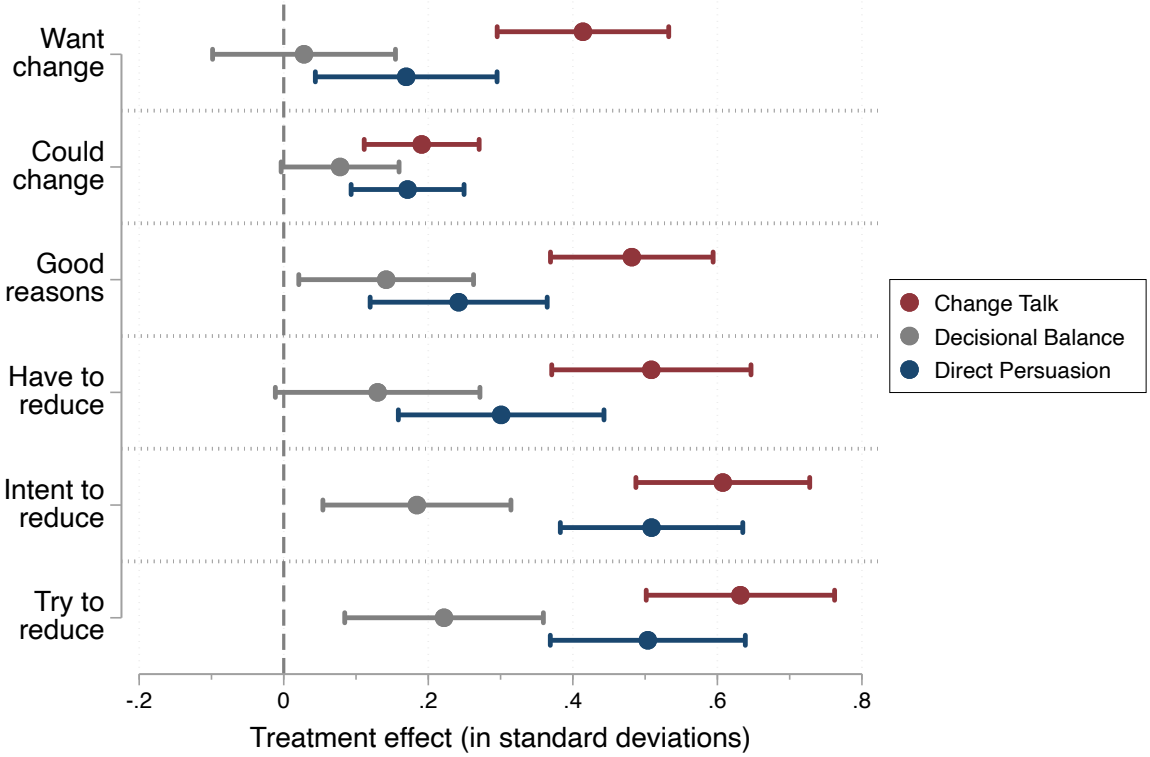
Note: This figure presents the average number of unique positive and the number of unique negative aspects of social media that are mentioned by participants during the conversation. We present summary statistics separately by treatment arm. Bars represent averages, and the error bars show the 95% confidence intervals.

Figure B.10: Strategies for reducing social media use mentioned during interviews



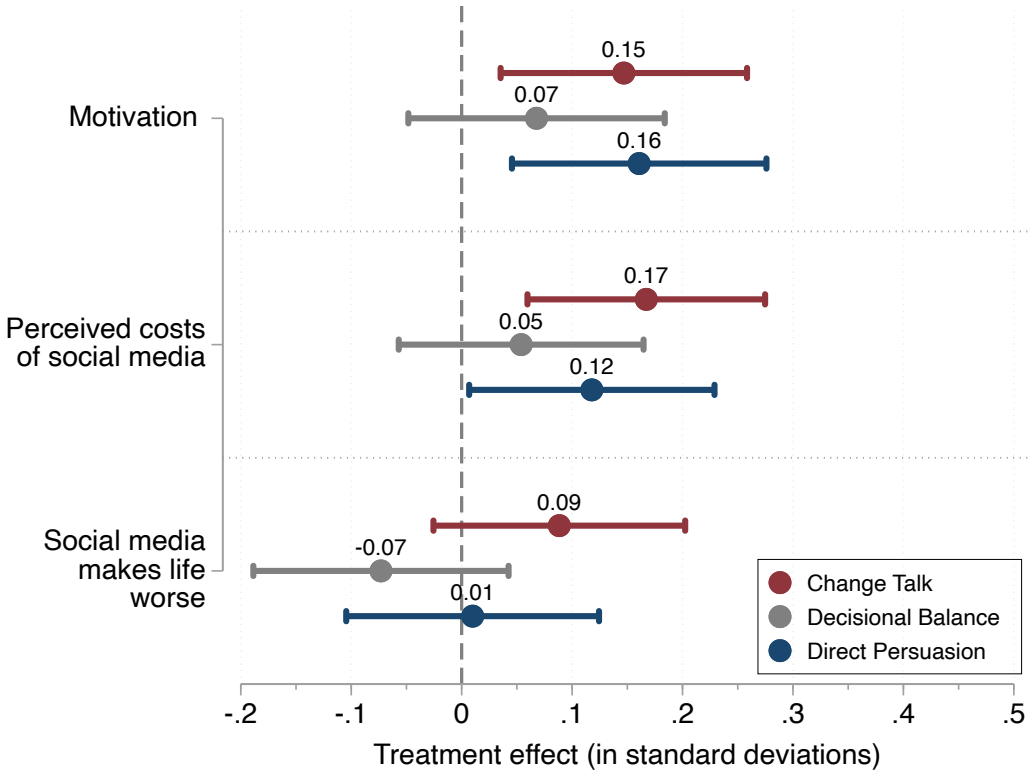
Note: This figure shows how often people mentioned a certain strategy to reduce their social media consumption during the interview. The different categories for the strategies are determined by us using an unsupervised topic analysis. We then use OpenAI's `GPT 5-nano` to classify the interview transcripts. Each interview transcript can contain more than one strategy.

Figure B.11: Treatment effects on subfacets of motivation to change



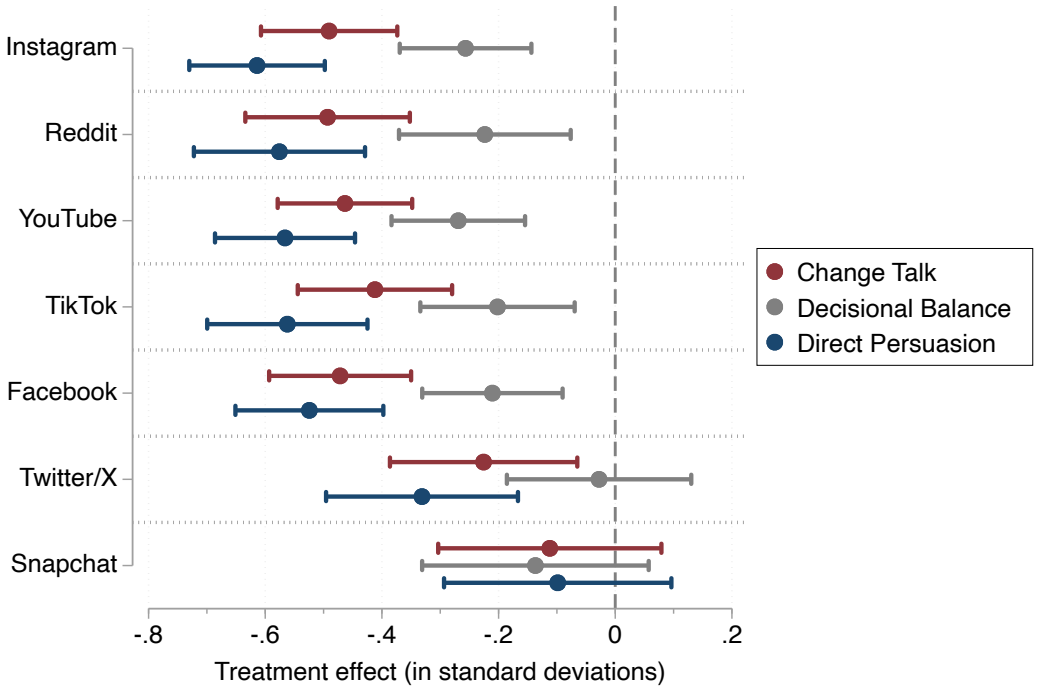
Note: This figure reports treatment effect estimates using data from the main survey. "Change Talk", "Decisional Balance" and "Direct Persuasion" are treatment indicators, with the time use interview control group being the omitted category. The dependent variables in each regression are standardized to have a mean of zero and a standard deviation of one among control group respondents. The dependent variables capture agreement with six items from the Change Questionnaire (Amrhein et al., 2003): "I want to reduce my social media use" (Want Change), "I could reduce my social media use" (Could change), "I have good reasons to reduce my social media use" (Good reasons), "I have to reduce my social media use" (Have to reduce), "I intend to reduce my social media use" (Intent to reduce), and "I am trying to reduce my social media use" (Try to reduce). Agreement is elicited on a 5-point Likert scale from "strongly disagree" to "strongly agree". 95% confidence intervals derived from robust standard errors are shown.

Figure B.12: Persistence of treatment effects on motivation and cost-benefit perceptions



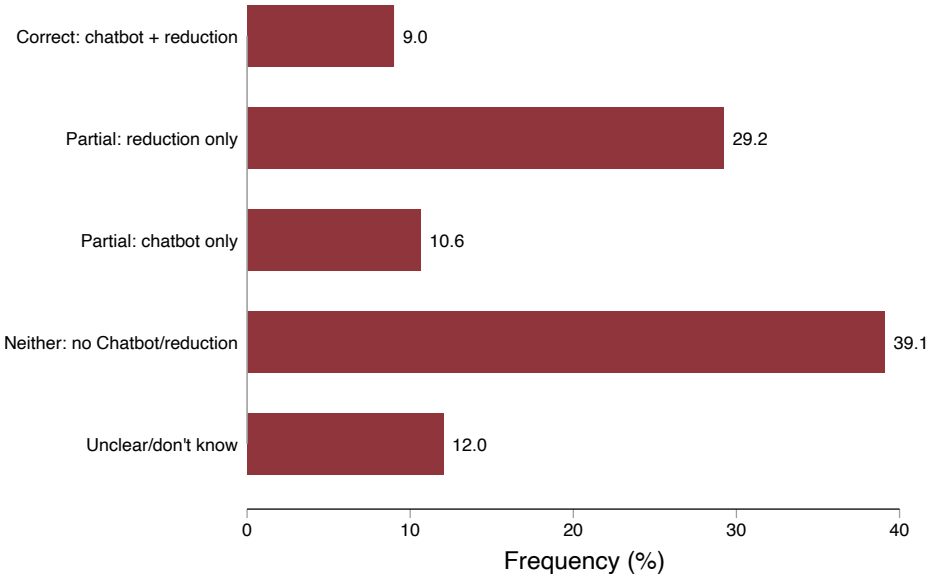
Note: This figure reports treatment effect estimates on motivation to reduce social media time as well as the cost-benefit perceptions using data from the follow-up survey. "Change Talk", "Decisional Balance" and "Direct Persuasion" are treatment indicators, with the time use interview control group being the omitted category. The dependent variables in each regression are standardized to have a mean of zero and a standard deviation of one in the control group. 95% confidence intervals derived from robust standard errors are shown.

Figure B.13: Treatment effects on predicted use of different social media apps



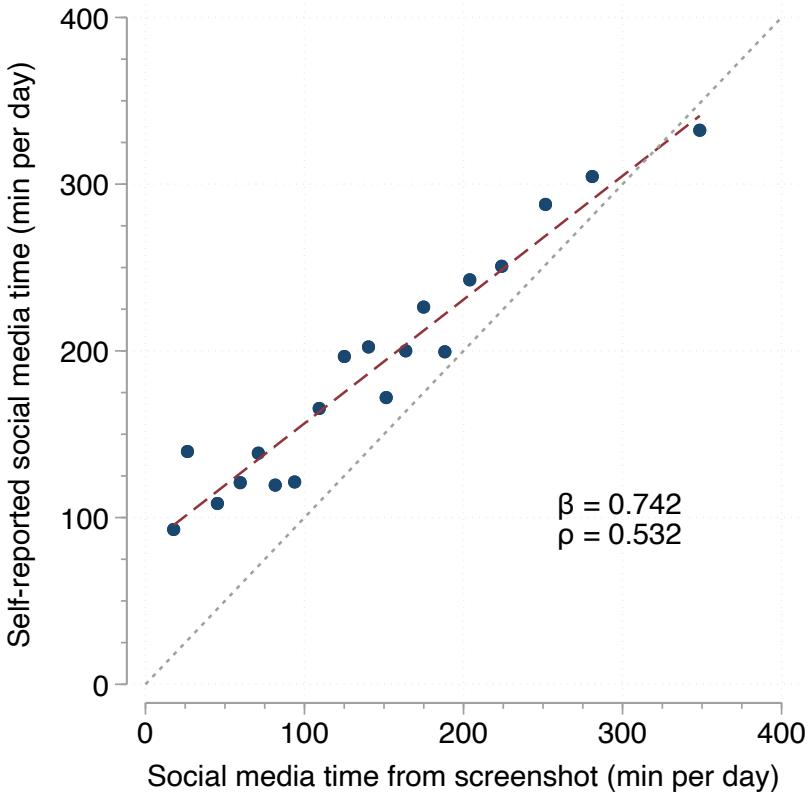
Note: This figure reports treatment effect estimates using data from the main survey. "Change Talk", "Decisional Balance" and "Direct Persuasion" are treatment indicators, with the time use interview control group being the omitted category. The dependent variables in each regression are standardized to have a mean of zero and a standard deviation of one among control group respondents. The dependent variable are derived from participants' responses to the question: "Compared to the past four weeks, how do you expect your time spent on each of the following apps to change over the next four weeks?" for TikTok, Instagram, Snapchat, Facebook, YouTube, Reddit, and Twitter/X. Responses are measured on a 6-point scale (much less; somewhat less; about the same; somewhat more; much more; I do not use this app) and are coded so that higher values indicate predicted increase in app use. For this analysis, we exclude participants that do not use an app. 95% confidence intervals derived from robust standard errors are shown.

Figure B.14: Categorization of answers given to the question about study purpose



Note: This figure shows a categorization of open-text answers participants give in our main survey when asked about the purpose of the study. We use OpenAI's `GPT-5.1-nano` to classify answers into one of the five categories shown in the graph. The categories are: (i) *Correct: chatbot + reduction* (e.g. "Gage whether people are interested in using social media less and whether they will trust AI models to help them do so", (ii) *Partial: reduction only* (e.g. "To help people reduce their social media time", (iii) *Partial: chatbot only* (e.g. "Training Chatbot.") (iv) *Neither: no Chatbot/reduction* (e.g. "Market research for an app", (v) *Unclear/don't know* (e.g. "No idea").

Figure B.15: Validation of self-reported social media use



Note: This figure presents a binscatter plot of self-reported average daily social media time over the previous two weeks against social media times measured from screenshot uploads of the iOS Screen Time app for the subsample of iPhone users with Screen Time activated on their iPhone. To derive this estimate, we extract time spent on apps in Apple's *Social* category, averaging values from calendar week 52 of 2025 and calendar week 1 of 2026. This category comprises the majority of apps relevant to our intervention (Instagram, TikTok, Snapchat, and Facebook). It also includes time spent on messaging services like WhatsApp, but excludes apps like YouTube and Reddit. We winsorize screen time at the 5th and 95th percentile. We report the bivariate regression coefficient (β) as well as the correlation between the two measures of social media time.

C Additional Evidence

C.1 Validation of the LLM measurements

This section provides additional details on the validation of our LLM-based pipeline for measuring MI treatment fidelity based on the MITI 4.2.1 coding manual (Moyers et al., 2014).

We use the following prompt template to score our motivational interview transcripts along the four global scores of the MITI 4.2.1 coding manual:

> # Task Overview
> You are an expert in evaluating motivational interviewing (MI) sessions. Your task is to rate a clinical session transcript based on a global component–{component_name}–from the Motivational Interviewing Treatment Integrity (MITI) 4.2.1 coding system.
>
> # Change Goal
> The session's change goal is to help the client reduce their weekly mobile phone screen time.
>
> # Coding Manual: {component_name}
> {coding_instructions}
>
> # Session Transcript
> {transcript}
>
> # Evaluation Instructions
> - Global components are assessed on a five-point Likert scale, with a minimum of '1' and a maximum of '5'. You assume a default score of '3' and move up or down as indicated in the Coding Manual. A '3' may also reflect mixed practice. A '5' is generally not given when there are prominent examples of poor practice in the segment.
> - Evaluate the clinician's performance using the Coding Scale (1-5) and return your score as an integer between 1 and 5.
> - If in doubt, assign the lower score if you are uncertain between two scores.
> - After determining your score, justify your choice in 1-2 sentences by referring to the main reason for choosing the score.
>
> # Output Format
> {{"score": <integer from 1 to 5>, "justification": "<1-2 sentence justification>"}}

We replace `{transcript}` with the full interview transcript, `{coding_instructions}` with the full verbatim instructions from the section in the MITI 4.2.1 that defines and explains how to evaluate the global component `{component_name}`. We use `gpt-5.1` for annotation, but `gpt-4.1-nano` delivers quantitatively similar results.

Table C.1 reports the results from validating LLM-assigned scores against a human ground truth derived from the 14 annotated transcripts that are part of the official training materials to

demonstrate both high and low quality implementations of MI principles encoded by the MITI 4.2.1 coding manual.[1]

Panel A presents the mean difference (bias) between LLM-assigned and human-assigned score, with negative values indicating that the LLM is more conservative than humans in assigning high scores. For each score, which is measured on a 5-point scale from 1 (low) to 5 (high), we see that LLM-assigned scores are close to human scores on average, with the LLM being slightly more conservative than human annotators, with a mean bias of -0.24 across all four scores. We thus interpret the LLM-based evaluation of our motivational interviews as a conservative bound of MI quality.

The final column presents bivariate correlation between LLM and human scores. Pooling across all four global scores, we obtain a global correlation of 0.72. Given the inherent subjectivity of the MITI 4.2.1 coding manual, we interpret this correlation as being sufficiently high. When examining the individual scores, we find that the LLM-based evaluation works best for the partnership (correlation of 0.89) and empathy scores (0.75).

In Panel B, we present analogous results for the so-called "behavioral counts" from the MITI 4.2.1 inventory. The behavioral counts are obtained by scoring each individual turn by the interviewer along a set of behavioral categories. For example, statements by the interviewer can be assigned codes such as "Question", "Advice without permission" or "reflection". These behavioral counts are then aggregate into interview-level counts of "MI-adherent" and "MI non-adherent" behaviors, as well as other summary statistics such as the ratio of reflections to questions of the share of complex reflections. Overall, correlations are high across nearly all categories, with the exception of the share of complex reflections.

To better understand this dimension, we additionally compute the correlation between whether the LLM and the human coder label an utterance as a reflection (regardless of whether it is simple or complex). This correlation is substantially higher. This pattern suggests that distinguishing between simple and complex reflections is currently challenging for LLMs. For this reason, we do not emphasize the complex-reflection share in the main analysis.

C.2 Scaling questions

Conversations in the two MI and the *Direct Persuasion* treatment arm all included two standardized scaling questions that ask participants to rate (i) how important it is for them to reduce their social media use and (ii) how confident they are that they can reduce their social media use on an 11-point scale. Figure B.5 presents CDF plots for both scores separately by treatment. Panel A shows that both *Change Talk* and *Direct Persuasion* shift the distribution of importance scores to the right relative to *Decisional Balance*, with the largest differences in the upper tail. *Change Talk* substantially increases the share of respondents reporting very high importance

[1] Available at `https://casaa.unm.edu/tools/miti.html`

Table C.1: Validation of LLM measurement for the MITI 4.2.1 coding manual

Score Category	**Bias**	**Correlation**
A. Global Metrics		
Global	-0.24	0.72
Partnership	-0.10	0.89
Cultivating Change Talk	-0.75	0.66
Empathy	0.15	0.75
Softening Sustain Talk	-0.25	0.45
B. Behavioral Counts		
Share Complex Reflections	-0.20	-0.01
Reflection-to-Question Ratio	-0.65	0.83
Total MI-Adherent Behavior	0.43	0.81
Total MI Non-Adherent Behavior	-0.43	0.84
Reflection Correlation		0.72

Note: Panel A reports the bias (mean difference between LLM and human assigned scores) and the correlation for the four global scores of the MITI 4.2.1 coding manual on a set of 14 fully annotated motivational interviews that are part of the official training material for demonstrating both high- and low-quality implementations of MI principles. We present statistics separately for each of the four global scores and jointly after pooling all scores ("Global"). Panel B reports summary statistics for the validation of LLM-based measurements of the "behavioral counts" from the MITI 4.2.1.

scores of 8 or above, consistent with its relatively stronger focus on inducing change talk. Panel B reports analogous patterns for confidence. Here, we find that participants in *Direct Persuasion* report the highest confidence scores, followed by *Decisional Balance* and *Change Talk*. Note that scaling questions are an MI technique to *increase* perceived importance and confidence, not a measurement of the importance and confidence in behavioral change as an outcome of participating in an MI. We therefore do not causally interpret these differences as reflecting underlying differences in the actual effectiveness of these treatments to motivate participants, which we separately measure after the interviews as one of our main outcomes and return to in Section 4.

D Example Interviews

Here, we present a full sample interview for each treatment.

D.1 Change Talk

Interviewer: Hi! I would like to talk with you about social media, such as Facebook, TikTok, and Instagram. What's the first thing that comes to mind when you think about your social media habits?

Participant: I binge social media, so I'll generally use most apps quite sparsely, but then when I do use social media apps, I'm on them for hours at a time, scrolling endlessly until I give up

Interviewer: It sounds like it's not constant for you, but when you do get on social media it can pull you in for hours and feels hard to stop. What are the downsides you notice when one of those long scrolling sessions happens, and what makes you wish your social media use were different?

Participant: I feel a bit tired and scrambled after a while. I'll know I'm getting bored but there isn't a better alternative, so I keep going, and then suddenly it's late and I've wasted my time

Interviewer: After those long scrolling sessions you end up feeling tired and scrambled, and it hits you that a lot of time slipped away. When do you notice those effects showing up the most in your day-to-day life?

Participant: In the evening or weekend mornings when I've scrolled after just waking up

Interviewer: Those evening and weekend-morning scrolls leave you feeling tired and scrambled, and then you realize a chunk of time is gone. When that happens, how does it get in the way of the kind of evening or morning—and the kind of person—you want to be?

Participant: I would love to be a morning person but I'll wake up with my phone, scroll until I'm bored and then it's suddenly early afternoon

Interviewer: You're noticing that those morning scrolls don't really match with the morning-person version of you you'd like to be, and that feels like time slipping away from what matters to you. If your social media use were more in line with the kind of person you want to be, what would a good weekend morning look like from the moment you wake up?

Participant: Wake up, make a cup of tea and read, do and chores and settle into bed with a hobby or a tv show to watch

Interviewer: Ok. Right now, when you pick up social media—especially on weekend mornings or evenings—it can turn into hours of scrolling that leaves you feeling tired and scrambled, and then suddenly half the day is gone. You'd really like your mornings to start with tea, reading, and getting some simple chores done, so you can relax into a hobby or a show feeling more intentional and like the morning-person you want to be. Did I get that right?

Participant: Yes!

Interviewer: You've got a clear picture of the kind of mornings and evenings you want—tea, reading, a few chores, then relaxing on purpose. On a scale from 0 to 10, where 0 means 'not at all important' and 10 means 'extremely important', how important is it to reduce the time you spend on social media?

Participant: 8

Interviewer: An 8 tells me this really matters to you and fits with the mornings and evenings you're aiming for—more tea, reading, and intentional relaxing. Why is it a 8 and not a lower number like zero?

Participant: It feels like wasted time - I don't use it mostly for actual socialising, just scrolling other people's posts. I could use the time to be more intentional and physical

Interviewer: You're really seeing that a lot of your social media time doesn't even feel social—it feels like time that gets taken from being intentional and more physically engaged in your day. If you were to reduce your social media use, what do you imagine might be the best or most important changes for you?

Participant: More energy, more time to spend doing useful things

Interviewer: You're wanting more energy and more time for useful, intentional things, and reducing social media feels like a real step toward that. On a scale from 0 to 10, where 0 means 'not at all confident' and 10 means 'extremely confident', how confident are you that you could follow through if you decided to reduce your social media time?

Participant: 5

Interviewer: Why is it a 5 and not lower number like zero?

Participant: I could probably do it if i wanted to - for example, I rarely even use my phone while at work

Interviewer: A 5 says part of you already believes you can do this, and you've seen proof of it—like how you rarely use your phone while you're at work. You can switch into a mode where social media just isn't part of what you're doing. What is it about work days—your routines, environment, or rules you follow—that you could borrow and apply to your mornings or evenings to make reducing social media feel more doable?

Participant: I need to be busy doing something else

Interviewer: That 5 in confidence makes sense—you've already shown you can keep social media out of the picture when you're engaged and busy, like you are at work, which says you do have some real control and ability here. Thinking back, what's a time in the past when you successfully changed a habit or followed through on a difficult change, and what did you do that helped it work?

Participant: I'm not too sure? I tend to give up on things. I've done sober months successfully just by saying no

Interviewer: Even with some doubts, you've shown you can follow through when you make a clear decision—like your sober months—and you already know being busy with something else helps you stay off social media. It also really matters to you to have more energy and use your time more intentionally. What ideas do you already have that feel realistic for reducing your social media time, especially on weekend mornings or in the evenings?

Participant: I need to pick up a hobby or a good book

Interviewer: You're already leaning toward a positive replacement—having a hobby or a good book ready so your time feels intentional instead of getting pulled into scrolling. That fits with the morning-person routine you described. If you decided to try this just as an experiment, what exactly would you do on a weekend morning—starting when you wake up—and how long would you want to keep social media off your phone before you check it (if at all)?

Participant: I'd want to respond to texts but not social media, so I'd send my boyfriend a good morning text and then leave my phone alone for a few hours

Interviewer: You've got a clear, doable plan: stay connected for texts, send your boyfriend a quick good-morning message, and then protect a few hours so your morning can be more intentional and energizing. As you wrap up, what feels like the most important takeaway or realization for you from this conversation about changing your social media use?

Participant: I need to spend my time better

Interviewer: You've noticed your social media use comes in long binges, especially on weekend mornings and evenings, and afterward you feel tired and scrambled and like the time has slipped away. You want to be more intentional—starting weekends with tea, reading, a few chores, and then relaxing on purpose—and you've identified a realistic experiment: send a quick good-morning text, then leave social media alone for a few hours while you do a hobby or read. That's the end of our interview. Thank you for your time and for being so open and thoughtful about what you want to change.

D.2 Decisional Balance

Interviewer: Hi! I would like to talk with you about social media, such as Facebook, TikTok, and Instagram. What's the first thing that comes to mind when you think about your social media habits?

Participant: I feel like I spend far too much time doom scrolling or watching YouTube Shorts.

Interviewer: Sounds like a lot of your time on social media can turn into doom scrolling and YouTube Shorts without you meaning to. What do you enjoy or find helpful about social media, or what would you miss if you spent less time on it?

Participant: I like the connection to friends and family that I don't see that much anymore on Facebook or Instagram. I also like the entertainment I get from YouTube. There are some streamers I've been watching for a long time, and I always feel happy watching them—it feels like it improves my wellbeing.

Interviewer: Staying connected with friends and family you don't see much, and having familiar YouTube streamers that reliably lift your mood, sound like real bright spots. What do those parts of social media add to your day when they're working well for you?

Participant: It just gives me a boost. When I've had a particularly bad day at work, sitting on the sofa watching YouTube or scrolling helps me disconnect from the real world and find some calm. It only becomes a problem when I look at the time and realize an hour has passed and I feel like I've wasted the day—even if it helped my mood.

Interviewer: Social media can really help you decompress after a rough workday and improve your mood, and at the same time noticing that an hour has disappeared can leave you feeling like you've lost valuable time.

What are some of the less-good parts of your social media use for you, especially in how it affects your time or the things you want to get done?

Participant: Once I start using social media, I find it hard to stop. I'll look at the clock and not realize that I've wasted a big chunk of the day when I have a million things to do in the real world. I almost need a timer to help me control how long I stay on it.

Interviewer: Once you get on social media it's hard to stop, and then a big chunk of time disappears when you had other things you meant to do—almost like you'd need an external timer to contain it. When do you notice this pattern most strongly, and what happens to the rest of your day when it does?

Participant: I notice it most strongly after a long day at work when I need to decompress and don't feel like doing anything sociable in real life. I choose to disconnect, mostly at night. During the day I can control it much better.

Interviewer: After a long workday, social media sounds like a dependable way to switch off and get a mood boost when you don't feel up for being sociable, and at the same time nighttime is when it's easiest for it to run longer than you mean to.

How do those nights fit—or clash—with the kind of person you want to be, or with the values that matter most to you?

Participant: They clash with the person I want to be. I want to be more in the moment, especially with my husband. Sometimes we'll both be sitting on the same sofa looking through our phones and not actually talking to each other for most of the night if we've both had a tough day.

Interviewer: On the one hand, social media gives you a real boost after hard days and helps you feel calmer. On the other hand, once you start at night it's hard to stop, time disappears, and it pulls you away from being present—like when you and your husband end up on your phones instead of talking. Did I capture that accurately?

Participant: That sounds exactly right.

Interviewer: On a scale from 0 to 10, where 0 means *not at all important* and 10 means *extremely important*, how important is it to reduce the time you spend on social media?

Participant: 7.

Interviewer: Why is it a 7 and not a lower number like zero?

Participant: Because it's important to me that I spend less time on social media and connect more with real life.

Interviewer: Why is it a 7 and not a higher number like ten?

Participant: Because social media has its place and helps my mood and wellbeing on difficult days.

Interviewer: Social media really does something important for you—especially after tough workdays—and at the same time you're feeling pulled toward more real-life presence, particularly with your husband.

If you reduced your social media use a bit, what do you imagine might change for you, both the positives and anything you might miss or find difficult?

Participant: I think I might do more fun things with my husband and learn more about each other. I would find it hard because it's relaxing to just go on social media.

Interviewer: On a scale from 0 to 10, where 0 means *not at all confident* and 10 means *extremely confident*, how confident are you that you could follow through if you decided to reduce your social media time?

Participant: 6.

Interviewer: Why is it a 6 and not a lower number like zero?

Participant: Because I really want to do it, and so does my husband, so we might be able to help each other.

Interviewer: Why is it a 6 and not a higher number like ten?

Participant: Because I know how much I like social media, and I haven't reduced it yet, so it would be difficult to start now.

Interviewer: Part of you feels that you and your husband could support each other and make this work, and another part knows how much you enjoy social media and that starting feels hard. When you think back to another time something was difficult to start or stick with but you managed it anyway, what helped you keep going?

Participant: My husband helped me to keep going.

Interviewer: Having your husband in your corner has helped you through hard changes before. With no pressure to act and just exploring options, what small step might you consider to gently adjust your social media use?

Participant: Limiting my time watching YouTube, or scheduling certain days when I don't go on it at all so it's not every day.

Interviewer: You're considering either limiting YouTube time or having specific no–social media days while keeping the relaxing part available. What's one version of that you could picture trying?

Participant: No YouTube on the weekend, so I can still use it after a hard day at work when I need it most.

Interviewer: You're leaning toward protecting YouTube for when it really helps after work, while creating weekend space to be more present with your husband and real life. After talking through both sides today, what stands out most to you?

Participant: I didn't realize how much I wanted to cut down on the time I spend on YouTube before talking this through.

Interviewer: You've described how social media—especially YouTube—helps you decompress, lift your mood, and stay connected, particularly after tough workdays. At the same time, once you start at night it's hard to stop, time disappears, and it can pull you away from being present with your husband and the real life you value. Through this conversation, you realized you want to cut down on YouTube more than you expected, and you're considering a weekend no-YouTube boundary while keeping it available when you truly need it. Thank you for your openness—this interview is now complete.

D.3 Direct Persuasion

Interviewer: I would like to have a conversation with you about your social media habits.

To set the stage, here is a brief summary of what research tends to find. In several large studies, people who took a short break from social media or cut back sharply for a few weeks often reported feeling better overall and spending more time with friends and family offline. Other research also finds that very heavy daily use is linked with worse sleep and feeling more down or stressed. Taken together, this suggests many people may be better off using social media much less than they currently do, especially if it has been affecting their focus, sleep, or time with others.

How does this land with you—does any of it fit your own experience with social media?

Participant: Yes, I agree with this. Every time I have taken time off scrolling social media, I have enjoyed furthering my other hobbies and socialising with family and friends offline.

Interviewer: You've already noticed that when you stop scrolling, you naturally spend more time on hobbies and real-life connection. Given what studies tend to show, a serious cutback is worth testing—especially if social media is getting in the way of your focus, sleep, or time with others. Where in your day would cutting back make the biggest difference for you right now—work time, evenings, or time with family and friends—and why?

Participant: Time with family and friends, so I can focus on them more.

Interviewer: You want to cut back mainly so you can be more present with family and friends. Which one or two apps are your biggest time sinks, and when do you tend to open them—for example during work, in the evening, or whenever you have a spare minute?

Participant: Twitter and Instagram are my biggest time sinks. Anytime I have a spare minute, or get a new message, I use them.

Interviewer: Twitter and Instagram are grabbing you in all the in-between moments, especially when a new message pops up. This kind of frequent checking is exactly the pattern that tends to improve when people do a serious cutback for a few weeks. What is the biggest cost for you right now—losing focus at work, losing quality time with others, or something else—and how do you notice it?

Participant: Losing focus on the things I want to be doing, like reading or working.

Interviewer: Twitter and Instagram are pulling you out of reading and work, and it's costing you the focus you actually want. What do you genuinely get from social media that you'd miss if you used it much less?

Participant: Social time with friends through texting, and keeping in touch with what's happening in the world.

Interviewer: Staying connected with friends and keeping informed are important benefits. The goal isn't to lose those, but to reduce the constant checking that disrupts your focus, sleep, or time with others. If you kept the benefits but cut usage down to a small, intentional amount, what would you most want to gain back—focus, calm, or time with others?

Participant: Focus.

Interviewer: You've said Twitter and Instagram are stealing your focus from reading and work, and getting that focus back is what you want most right now. On a scale from 0 to 10, where 0 means *not at all important* and 10 means *extremely important*, how important is it for you to substantially reduce your social media time—at least for a few weeks—to see how you feel?

Participant: 5.

Interviewer: You rated the importance as a 5, which suggests you see some meaningful upside in cutting back. What specific, measurable cutback plan would you be willing to try for the next few weeks?

Participant: Reducing my time on Twitter to one hour a day.

Interviewer: You're willing to cap Twitter at one hour a day, which is a concrete start. However, that may be too mild to really restore your focus. Benchmarks that often work better are: (A) a complete break from all social media for at least two weeks, or (B) about 30 minutes total per day for at least three weeks. Do you want to commit to A, B, or a stricter version of your one-hour Twitter plan?

Participant: B seems like a good idea.

Interviewer: Good choice—30 minutes total per day is much more likely to give you your focus back. What one clear daily rule will you follow for the next three weeks?

Participant: I will use Twitter for a maximum of 30 minutes total per day.

Interviewer: Your rule is that you'll use Twitter for a maximum of 30 minutes total per day for the next three weeks. What is the main way you will enforce this so it doesn't rely on willpower—built-in limits, a blocker, deleting the app, or something else?

Participant: A blocker.

Interviewer: A blocker is a smart choice because it reduces the need to fight temptation in the moment. On a scale from 0 to 10, where 0 means *not at all confident* and 10 means *extremely confident*, how confident are you that you could follow through with this plan?

Participant: 7.

Interviewer: A 7 is solid. What do you think will be the main obstacle that could still derail you from sticking to 30 minutes a day?

Participant: A lack of willpower.

Interviewer: Willpower is often unreliable here. Pairing your blocker with a passcode-protected app timer or a stricter third-party blocker can help. What single first step will you take today to set up that enforcement?

Participant: Pairing my blocker with a passcode-protected app timer.

Interviewer: You want to be more present with family and friends, and you've committed to using Twitter for a maximum of 30 minutes total per day for the next three weeks, supported by a blocker and a passcode-protected app timer. Thank you for your time—this concludes our interview.

D.4 Control

Interviewer: Hi! In this interview I want to learn more about how you spend your time. Could you start by walking me through the practical steps of how you get your day started in the morning—from waking up until you're ready to begin your day?

Participant: Hello. I wake up at 6:15am. I get dressed, grab breakfast, and go out walking to do approximately 10,000 steps. I then come home, get a cup of tea, wake the kids, and get them ready for school. While they are eating breakfast, I grab a shower and get dressed again, ready for work. I then drop the kids at school at 8:45 and arrive at work at 9am, where I am ready to start my day properly.

Interviewer: You've got a really structured morning—up at 6:15, a long walk, then tea, getting the kids up and ready, a quick shower, and school drop-off before you're at work by 9. Is there anything else you usually fit into that morning window, like checking messages, packing lunches, or planning your day?

Participant: No, there's no time. My husband prepares lunches, or else the kids have school dinners.

Interviewer: It sounds like your mornings are pretty full already, and with lunches handled there isn't really any extra time to squeeze things in. Once you're at work around 9, how does your day usually flow from late morning into midday or early afternoon?

Participant: I arrive at work, switch on the computer, check my emails and messages, and then do a bit of work. My next cup of tea is around 10am, but I don't stop and go straight back to work until my actual break time at 11am. I visit clients or type up work all morning, with a short 15-minute break at 11am. Lunch is then at 1pm.

Interviewer: So your late morning is mostly focused work—emails first, a quick tea around 10, a short break at 11, then client visits or writing until lunch at 1. After work, how do you usually spend your evenings and wind down before bed?

Participant: With four kids, my evenings are hectic. Dinner, homework, a bit of TV, or taking them to their activities, then a bedtime snack and getting them to bed. I don't sit down until around 9pm most nights.

Interviewer: Your evenings sound really busy and kid-focused—dinner, homework, activities, a bit of TV, then snacks and bedtime, and you're not sitting down until around 9. Once you finally sit down, what do you usually do in that last stretch before sleep?

Participant: I watch TV for the most part and chat with my husband. That's about it.

Interviewer: When it comes to planning, do you usually plan your day in advance or figure it out as you go, and what does that look like for you?

Participant: My days are usually very much planned, but when I do have free space I love to go out on an adventure.

Interviewer: It sounds like most days you keep things pretty planned out, but when you do get free space you enjoy heading out on an adventure. Are there any specific habits or routines you rely on most days to keep everything running smoothly?

Participant: Just the general structure to get everything done.

Interviewer: Do your routines change across seasons—for example, summer versus winter—and if so, how?

Participant: No, not at all.

Interviewer: So even across summer and winter, your routine stays basically the same. What is the biggest difference season to season in how you spend your time?

Participant: I just have less energy in winter because there is less light.

Interviewer: Even though your schedule stays consistent across seasons, winter takes more out of you. How do your routines change on weekends, if at all, compared to weekdays?

Participant: They don't really change. We are a family who enjoy going out on long walks and day trips no matter the season.

Interviewer: You've described very consistent routines across weekdays, weekends, and seasons, with structure and family activities playing a central role. Is there anything else you would like to add or reflect on about how you spend your time day to day?

Participant: Not really. It's just so important to have that structure, given we are a large family, to make sure everyone gets what they need.

Interviewer: You described a highly structured day: up at 6:15 for a long walk, then getting the kids ready and to school before starting work at 9, with a busy morning of emails, client visits, and writing until lunch. Evenings are largely focused on your children—dinner, homework, activities, and bedtime—before you finally sit down around 9 to relax with your husband. Weekends follow a similar pattern, with family walks and day trips throughout the year. Thank you for sharing your routines and how structure helps your family function. This concludes the interview.

E Experimental Instructions

We provide the experimental instructions below. Text shown in [italics] is not shown to respondents. Response options are shown in square brackets. Page breaks are indicated.

E.1 Main survey

[Captcha and consent form]

[Page break]

Please follow the instructions in the audio to answer the question below. This is to identify bots, as they cannot interpret audio easily.
[Audio file with the statement: "Please enter the number 5."]

Your answer: _____

[Page break]

The next question is about the following problem. In questionnaires like ours, sometimes there are participants who do not carefully read the questions and just quickly click through the survey. This means that there are a lot of random answers which compromise the results of research studies. To show that you read our questions carefully, please choose both "Very strongly interested" and "Not at all interested" as your answer in this question.

How interested are you in sports?

- Very strongly interested
- Very interested
- A little bit interested
- Almost not interested
- Not at all interested

[Respondents that do not pass the above attention checks are screened out of the survey.]

[Page break]

What is your favorite month of the year? Please explain in about 15 to 30 words. _____

[We preregistered to ex-post exclude participants that write fewer than 20 characters or that write faster than 10 characters per second on this task.]

[Page break]

What is the operating system of your primary phone?
[Android, iOS (iPhone), Other/Don't know]

[Page break]

Your social media use

We will now turn to questions about your social media use. We are always referring to these apps and platforms: TikTok, Instagram, Snapchat, Facebook, YouTube, Reddit, and X/Twitter.

Which of the following apps are currently installed on your smartphone?
[TikTok, Instagram, Snapchat, Facebook, YouTube, Reddit, X/Twitter, None of the above]

[Page break]

On a typical day, how much total time do you spend on social media? (TikTok, Instagram, Snapchat, Facebook, YouTube, Reddit, and X/Twitter)

Daily total social media screen time: ____ hours ____ minutes

[Page break]

You just told us that you typically spend about 2 hours and 2 minutes per day on social media apps.

How much total time would you ideally like to spend on social media per day?

Ideal total social media screen time: ____ hours ____ minutes

[Page break]

On a typical day, how much time do you spend on each of these apps?

- TikTok
- Instagram
- Snapchat
- Facebook
- YouTube
- Reddit
- X/Twitter

[Response options: No time, 0-5 min, 5-15 min, 15-30 min, 30-60 min, 60-120 min, More than 120 min]

[Page break]

Thinking about your current use of the social media platforms below, do you use them more or less than you would ideally like?

- TikTok
- Instagram
- Snapchat
- Facebook
- YouTube
- Reddit
- X/Twitter

[Response options: I do not use this app, Much less than I would like, Somewhat less than I would like, About the right amount, Somewhat more than I would like, Much more than I would like]

[Page break]

How addicted do you feel towards the social media apps below?

- TikTok
- Instagram
- Snapchat
- Facebook
- YouTube
- Reddit
- X/Twitter

[Response options: I do not use this app, Not at all addicted, A little addicted, Somewhat addicted, Addicted, Very addicted]

[Page break]

Interview

On the next page, you'll participate in an interview with a chatbot about your [social media use / time use].

This interview is a very important part of the study. Please answer thoughtfully and take your time.

You have two options to answer the questions: You type the answers using your keyboard. You use the voice recording feature, see how it works in the video below.

[Short video explainer]

[Page break]

Interview

The interview will take on average 15 minutes to complete. Once you've answered all questions, you'll be able to continue with the survey.

(Interview interface)

[Page break]

Thank you for completing the interview. On the next few pages, we will ask you some questions about your views on social media. There are no right or wrong answers to these questions, we are simply interested in your own views.

[Page break]

How motivated are you to reduce your time on social media apps?

[Not at all motivated 0, 1, 2, ..., 9, Extremely motivated 10]

[Page break]

How much do you agree or disagree with each of the following statements?

- Spending too much time on social media makes it harder for me to focus on important tasks.
- Excessive social media use reduces the quality of my social interactions.
- Reducing my social media use would help me feel more present and engaged in my daily life.
- Using social media less would improve my sleep and overall well-being.

[Response options: Strongly disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Strongly agree]

[Page break]

How much do you agree or disagree with each of the following statements?

- I can always manage to solve difficult problems if I try hard enough.
- When I set my mind to something, I usually succeed.
- Even if others doubt me, I believe in my ability to achieve my goals.
- I can usually find ways to overcome obstacles.

[Response options: Strongly disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Strongly agree]

[Page break]

How much do you agree or disagree with each of the following statements?

- I often use social media for longer than I originally intended.
- Even when I try to limit my social media use, I find it difficult to stop.
- I sometimes feel frustrated with myself for not being able to control how much I use social media.

[Response options: Strongly disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Strongly agree]

[Page break]

How much do you agree or disagree with each of the following statements?

- I would be embarrassed if people knew how much time I actually spend on social media.
- It is important for me to think of myself as a person in control of my life with healthy social media habits.

[Response options: Strongly disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Strongly agree]

[Page break]

To what extent do you think your current social media use makes your life better or worse? [-5 (Makes my life worse), -4, ..., 0 (neutral), 1, ..., +5 (Makes my life better)]

[Page break]

How much total time would you ideally like to spend on social media per day?

Ideal total social media screen time: ____ hours ____ minutes

[Page break]

Compared to the past four weeks, how do you expect your time spent on each of the following apps to change over the next four weeks?

- TikTok
- Instagram
- Snapchat
- Facebook
- YouTube
- Reddit
- X/Twitter

[Response options: I do not use this app, Much less, Somewhat less, About the same time, Somewhat more, Much more]

[Page break]

How much time do you think you will actually spend on social media per day on average over the next four weeks?

Your predicted daily social media screen time over the next four weeks: ____ hours ____ minutes

[Page break]

How much do you agree or disagree with each of the following statements?

- I want to reduce my social media use
- I could reduce my social media use
- I have good reasons to reduce my social media use
- I have to reduce my social media use
- I intend to reduce my social media use
- I am trying to reduce my social media use

[Response options: Strongly disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Strongly agree]

[Page break]

We now want to ask you a few questions about the interview you just had.

[Page break]

For each statement, please select the option that best describes the behavior of the interviewer in the conversation you previously had in this survey.

- Help you to talk about changing your behavior
- Make you talk about something you didn't want to discuss
- Help you discuss your need to change your behavior
- Help you discuss the pros and cons of your behavior
- Argued with you to change your behavior
- Help you feel hopeful about changing your behavior
- Act as a partner in your behavior change
- Help you recognize the need to change your behavior
- Tell you what to do
- Help you feel confident in your ability to change your behavior
- Act as an authority on your life
- Make you feel pressured to change your behavior

[Response options: Not at all, A little, Sometimes, A great deal, Always]

[Page break]

Please think about the interview you had as part of this survey.

Indicate the extent to which you have felt the following emotions during the interview:

- Interested
- Excited
- Upset
- Guilty
- Irritated
- Ashamed
- Determined
- Encouraged

[Response options: Very slightly or not at all, A little, Moderately, Quite a bit, Extremely]

[Page break]

For me, thinking about reducing my social media use feels: [Very pleasant, Pleasant, Neutral, Unpleasant, Very unpleasant]

[Page break]

Do you currently use any apps that help you manage your social media use (by setting time limits or scheduled blocks on social media apps)?
[Yes, No]

[Page break]

How interested would you be in an app that helps you manage your social media screen time?
[Not at all interested, A little bit interested, Somewhat interested, Interested, Very interested, Extremely interested]

[Page break]

Freedom app

Freedom is an app and website blocker that helps you limit your use of social media and other distracting apps on your phone, tablet, and computer.

On the next page, you will see a brief introduction to how Freedom works, and then we will ask you some questions to understand how valuable you find this type of service.

We are not affiliated with the company offering the Freedom app, and this is not promotional material.

Are you familiar with the Freedom app?
[No, I've never heard of it; I've heard of it, but never used it; I've used it in the past, but I don't currently use it; Yes, I currently use it]

[Page break]

Manage your social media time with the Freedom app

Freedom is an app and website blocker that helps you limit your use of social media and other distracting apps – for example, Instagram, TikTok, Snapchat, and Facebook. You decide which apps and sites to block, and when to block them.

- Set timers for specific apps (for example, block Instagram and TikTok from 8 pm–11 pm or during work hours).
- Locked Mode makes it hard to stop a block early, for extra commitment.
- Works across devices (phone, tablet, computer) so the same social media apps are blocked everywhere.

On the next page, we will ask you how much you would value one-year access to Freedom Premium, which includes premium features such as unlimited sessions and devices, recurring schedules, and other tools to help you stick to your limits.

[Page break]

Your valuation of the Freedom app

You are about to make choices that can have real consequences for you. After the study, we will randomly select 100 participants. If you are selected, we will randomly pick one of the rows from the table below and implement the option you chose in that row. It is therefore in your best interest to answer these questions truthfully and accurately.

In this task, we would like to understand how much you value one-year access to Freedom Premium. For each row in the table below, please choose between two options:

- Option A: One-year access to Freedom Premium
- Option B: A bonus payment of the amount shown in that row, added to your Prolific account

If Option A is selected for you, we will send you a voucher code via Prolific. This code gives you one-year access to Freedom Premium. To redeem it, simply create an account at

https://freedom.to/trial and enter the code when prompted. No credit card is required at any point, and no payment information will be collected or stored.

Which option do you prefer?
[Multiple price list: Option A is always "1 year access to Freedom" and Option B is "$[X] bonus" with X$\in [0, 1, \ldots, 9, 10, 15, 20]$]

[Page break]

On the previous page, your answers implied that you would prefer one-year access to Freedom Premium over a $[X] bonus, but that you would prefer a $[Y] bonus over a one-year access to the Freedom Premium.

Is this correct? If not, you can adjust your answer on the next page. [Yes, this is correct; No, I would also prefer one year access to the Freedom premium plan over a $Y bonus.; No, I would also prefer a $X bonus over a one year access to the Freedom premium plan.]

[If one of the "No" options is selected, we repeat the elicitation on the previous page.]

[Page break]

How much do you trust AI technology, such as large language models like ChatGPT, in general?
[A great deal, A lot, A moderate amount, A little, None at all]

How often do you use ChatGPT or other large language models?
[Every day, Multiple times per week, A few times a month, Once a month, Never]

[Page break]

In which country do you currently reside?
[United States, United Kingdom]

[Page break]

What is your gender?
[Male, Female, Other / Prefer not to say]

What is your age? ____

Which of the following best describes your race or ethnicity? [For US respondents: African American/Black; Asian/Asian American; Caucasian/White; Native American, Inuit or Aleut; Native Hawaiian/Pacific Islander; Other; For UK respondents: Black, Black British, Caribbean or African; Asian/Asian British; Caucasian/White; Mixed or multiple ethnic groups; Other ethnic group]

Are you of Hispanic, Latino, or Spanish origin?
[Yes, No]

What is your current employment status? [Full time employee, Part-time employee, Self-employed or small business owner, Unemployed and looking for work, Student, Not in labor force (for example: retired or full-time parent]

What is the highest level of education you have completed? [For US respondents: Some high school or less; High school diploma or GED; Some college, but no degree; Associates or technical degree; Bachelor's degree; Graduate or professional degree (MA, MS, MBA, PhD, JD, MD, DDS etc.); For UK respondents: Some primary; Completed primary school; Some secondary; Completed secondary school; Vocational or similar; Some university but no degree; Bachelor's degree; Graduate or professional degree (MA, MS, MBA, PhD, JD, MD, DDS)]

What was your annual household income in 2024 in US dollars before taxes and deductions? Note: The household income is the total amount of money earned by every member of your household. [Less than 15,000; Between 15,000 and 25,000; Between 25,000 and 50,000; Between 50,000 and 75,000; Between 75,000 and 100,000; Between 100,000 and 150,000; Between 150,000 and 200,000; More than 200,000]

Which of the below best describes your political affiliation? [For US respondents: Strong Democrat, Weak Democrat, Independent, Weak Republican, Strong Republican; For UK respondents: Strong Labour, Weak Labour, Independent / No affiliation, Weak Conservative, Strong Conservative]

Who did you vote for in the [2024 UK general election / 2024 presidential election between Donald Trump and Kamala Harris], or did you not vote? [For US respondents: Donald Trump, Kamala Harris, Someone else, I did not vote; For UK respondents: Conservative Party, Labour Party, Liberal Democrats, Green Party, Scottish National Party (SNP), Plaid Cymru, Reform UK, UK Independence Party (UKIP), Other party (please specify), I did not vote, I was not eligible to vote]

[Page break]

What do you think was the purpose of this study? _____

[Page break]

Thank you for completing this survey!

If you have any feedback or comments, please let us know in the text box below. _____

E.2 Follow-up survey

A few weeks ago, you told us that you typically spend about [X] hours and [Y] minutes per day on social media apps (TikTok, Instagram, Snapchat, Facebook, YouTube, Reddit, and X/Twitter).

How motivated are you to spend less time than this on social media apps this year?
[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, Extremely motivated 10]

[Page break]

How much do you agree or disagree with each of the following statements?

- Spending too much time on social media makes it harder for me to focus on important tasks.
- Excessive social media use reduces the quality of my social interactions.
- Spending too much time on social media makes it more difficult to be present and engaged in my daily life.
- Spending too much time on social media worsens my sleep and overall well-being.

[Response options: Strongly disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Strongly agree]

[Page break]

To what extent do you think that social media makes your life better or worse?
[-5 (Makes my life worse), -4, -3, -2, -1, 0 (neutral), +1, +2, +3, +4, +5 (Makes my life better)]

[Page break]

Over the last two weeks, how much time have you spent on social media per day, on average? (TikTok, Instagram, Snapchat, Facebook, YouTube, Reddit, X/Twitter)

We recommend that you look up your previous screen time on your phone before answering (use the iPhone "Screen Time" app or the Android "Digital Wellbeing" app).

Average daily social media screen time: ____ hours ____ minutes

[Page break]

Which, if any, steps have you taken to reduce your time on social media apps over the last two weeks? Select all that apply.

- I used built-in phone settings (e.g., Screen Time/Digital Wellbeing) to track or restrict my social media use
- I used a third-party app blocker to restrict access at certain times
- I turned off or muted some or all social media app notifications
- I made social media harder to access without deleting apps (e.g., logged out, hid/moved apps, removed home-screen shortcuts)
- I deleted or uninstalled one or more social media apps (temporarily or permanently), or deactivated an account
- I set specific rules/goals for myself about social media use (e.g., only once per day, no phone at bedtime)

- I put my phone out of reach to reduce temptation (e.g., another room)
- I reduced my overall phone use to avoid being tempted by social media apps
- I changed what I see on social media to reduce temptation (e.g., unfollowed/muted accounts)
- I replaced social media with other activities or routines (e.g., reading, hobbies, exercise)
- I asked someone else to help (e.g., accountability partner, family/friend, someone else set the passcode/lock)
- I mostly relied on willpower/discipline to resist the urge to use social media
- Other (please specify): _____
- I haven't taken any steps to reduce my social media time

[Page break]

In this part, we will ask you to upload a screenshot of your social media screen time last week [correct date range shown here].

We will pay you a bonus of $1 for uploading this screenshot.

Here are step-by-step instructions on how to access your screen time:

1. Go to Settings and tap on Screen Time
2. Then tap on See all App & Website Activity
3. Make sure you have selected Week
4. Now, swipe to see Last Week's Average
5. Now, swipe down and tap on "Show Categories"
6. Tap on the "Social" category
7. Now, you can see information with your screentime on your social apps. Please upload this screenshot.

[Visual step-by-step instructions with iPhone screenshots are displayed]

Please upload a screenshot of your screen time over the last week below to receive the $1 bonus. The screenshot must include last week's average social media screen time and the full week chart including the time spent in different apps below the chart (as shown in step 7 above).
[File upload field]

[Page break]

Screen time for the week before last week

Now, we ask you about your social media screen time for the week one week before last week [correct date range shown here] as recorded by your phone.

We will pay you an additional bonus of $1 for sharing a screenshot of your screen time from your phone to verify this information.

Here are step by step instructions how to access the screenshot:

1. If you are still on the screen showing last week's social media screentime, go back
2. On the main screen showing your total screentime, swipe until you see the week [correct date range shown here]
3. Swipe down and tap on "Show Categories"

4. Tap on the "Social" category
5. Now, you can see information with your screen time on your social apps over the week [correct date range shown here]. Please upload this screenshot.

[Visual step-by-step instructions with iPhone screenshots are displayed]

[Note box: If you got confused somewhere, you can see the full instructions from the start at the bottom of the page again!]

Please upload a screenshot of your screen time over the week of [correct date range shown here] to receive another $1 bonus. The screenshot must include the average screen time and the full week chart including the time spent on different apps below the chart (as shown above in step 7).

[Expandable link:] Click here to see the full instructions again on how to retrieve the screenshot from the week before last week if you start from the home screen.
[File upload field]

[Page break]

Screen time for the week two weeks before last week

Now, we ask you about your social media screen time for the week two weeks before last week [correct date range shown here] as recorded by your phone.

We will pay you an additional bonus of $1 for sharing a screenshot of your screen time from your phone to verify this information.

Here are step by step instructions how to access the screenshot:

1. If you are still on the screen showing last week's social media screentime, go back
2. On the main screen showing your total screentime, swipe until you see the week [correct date range shown here]
3. Swipe down and tap on "Show Categories"
4. Tap on the "Social" category
5. Now, you can see information with your screen time on your social apps. Please upload this screenshot.

[Visual step-by-step instructions with iPhone screenshots are displayed]

[Note box: If you got confused somewhere, you can see the full instructions from the start at the bottom of the page again!]

Please upload a screenshot of your screen time over the week of [correct date range shown here] to receive another $1 bonus. The screenshot must include the average screen time and the full week chart including the time spent in different app-categories below the chart (as shown above in step 7).

[Expandable link:] Click here to see the full instructions again on how to retrieve the screenshot from the week two weeks before last week if you start from the home screen.
[File upload field]