Optimizing Social Utility in Sequential Experiments

Ander Artola Velasco [1] **Stratis Tsirtsis** [2] **Manuel Gomez Rodriguez** [1]

Abstract

Regulatory approval of products in high-stakes domains such as drug development requires statistical evidence of safety and efficacy through large-scale randomized controlled trials. However, the high financial cost of these trials may deter developers who lack absolute certainty in their product's efficacy, ultimately stifling the development of 'moonshot' products that could offer high social utility. To address this inefficiency, in this paper, we introduce a statistical protocol for experimentation where the product developer (the agent) conducts a randomized controlled trial sequentially and the regulator (the principal) partially subsidizes its cost. By modeling the protocol using a belief Markov decision process, we show that the agent's optimal strategy can be found efficiently using dynamic programming. Further, we show that the social utility is a piecewise linear and convex function over the subsidy level the principal selects, and thus the socially optimal subsidy can also be found efficiently using divide-and-conquer. Simulation experiments using publicly available data on antibiotic development and approval demonstrate that our statistical protocol can be used to increase social utility by more than 35% relative to standard, non-sequential protocols.

1. Introduction

Access to markets in high-stakes domains, such as drug development, is strictly governed by regulatory bodies to ensure that new products meet rigorous safety and efficacy standards (Gieringer, 1985; U.S. Food and Drug Administration, 2019). Randomized controlled trials (RCTs) serve as one of the primary mechanisms for access control, requiring developers to gather sufficient statistical evidence to prove that a product is safe and effective (O'Connor, 2010; Janiaud et al., 2021; Farina et al., 2021).

However, the high financial costs of RCTs can stifle the development of 'moonshot' products that could offer high social utility, as developers may hesitate to proceed if they lack absolute certainty in their product's efficacy (Detsky, 1989; Frantz, 2003; Martin et al., 2017). To avoid missing such opportunities, calls have emerged for adaptive RCTs (Brown et al., 2009; Mahajan & Gupta, 2010; U.S. Food and Drug Administration, 2026) and targeted subsidies (National Institutes of Health, 2026a; U.S. Food and Drug Administration, 2024; Deutsche Forschungsgemeinschaft, 2024; European and Developing Countries Clinical Trials Partnership, 2025), particularly in the context of orphan and rare diseases where small patient populations often discourage private investment.

Yet, the effectiveness of such interventions depends on how developers strategically respond to them. In this context, a recent line of work (Bates et al., 2022; Shi et al., 2024; Hossain et al., 2025) has argued that the regulatory approval of products is best modeled as a principal-agent game (Grossman & Hart, 1992), where the regulator (the principal) designs an approval protocol and a hypothesis testing rule to incentivize the product developer (the agent) to act in a way that aligns with the principal's interests. In this paper, we extend this line of work to consider a setting in which the RCTs are conducted sequentially and are (partially) subsidized by the principal.

Our contributions. We introduce a statistical protocol for experimentation where the product developer (the agent) conducts an RCT sequentially and the regulator (the principal) partially subsidizes its cost. At each step of the protocol, the agent and principal update their beliefs about the product's effectiveness based on the latest experimental outcome. If there is sufficient statistical evidence to reject the null hypothesis (Ramdas et al., 2023), the principal approves the product; otherwise, they request that the agent gather more evidence, and the agent may either proceed with the experiment or terminate without approval. By modeling the protocol using a belief Markov decision process (Kaelbling et al., 1998), we show that the agent's optimal strategy can be found efficiently using dynamic programming. Further, we show that the social utility is a piecewise linear and

[1]Max Planck Institute for Software Systems, Kaiserslautern, Germany [2]Hasso Plattner Institute, Potsdam, Germany. Correspondence to: Ander Artola Velasco <avelasco@mpi-sws.org>.

ICML 2026 Workshop on Hypothesis Testing, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

convex function over the subsidy level the principal selects, and thus the socially optimal subsidy can also be found efficiently via divide-and-conquer.

To validate our statistical protocol, we conduct simulation experiments using publicly available data on antibiotic development and approval. The code for our experiments is available at `https://github.com/Human-Centric-Machine-Learning/strategic-experiments`. The results show that our protocol can be used to increase social utility by more than 35% relative to standard, non-sequential protocols.

Further related work. Our work builds upon further related work on the economic aspects of statistical testing, sequential hypothesis testing, and Bayesian experimental design.

A recent and closely related line of work studies the economic and strategic incentives arising in regulatory approval processes (Tetenov, 2016; Bates et al., 2022; McClellan, 2022; Shi et al., 2024; Hossain et al., 2025). Therein, Shi et al. (Shi et al., 2024) and Hossain et al. (Hossain et al., 2025) focus on determining the optimal hypothesis test that the principal can use to control false positives and false negatives, while Tetenov (Tetenov, 2016) and Bates et al. (Bates et al., 2022) study, respectively, the design of approval and payment rules to disincentivize agents who know their product is ineffective from participating in the approval process. Relatedly, McClellan (McClellan, 2022) considers a setting in which the principal designs approval rules to encourage agent participation without monetary transfers, *e.g.*, by lowering approval standards if previous experiments were not successful. In contrast to these works, we study how the principal can optimally subsidize the agent to increase social utility in a scenario where both are uncertain about the effectiveness of the product.

Within the hypothesis testing literature, our work draws on an active line of research on anytime-valid statistical inference using e-values (Ramdas et al., 2023; Grünwald et al., 2024; Ramdas & Wang, 2025), which has been successfully applied to a wide range of statistical problems (Waudby-Smith & Ramdas, 2022; Xu & Ramdas, 2024; Shin et al., 2024; Shekhar & Ramdas, 2024; Waudby-Smith et al., 2025; Chugg et al., 2026; Bates et al., 2022; Gauthier et al., 2025; Huang et al., 2026; Velasco et al., 2026; Dhillon et al., 2026). Most closely related to ours is the work of Bates et al. (Bates et al., 2022), who also uses e-values in the context of RCTs. However, their focus is on designing contracts that disincentivize agents with ineffective products from participating in the approval process. We instead use e-values to design an approval protocol that provides sequential error guarantees to the principal.

Our work also connects to the broad literature on Bayesian experimental design (Robbins, 1970; Peskir & Shiryaev, 2006; Powell & Ryzhov, 2012; Ghavamzadeh et al., 2015; Cho et al., 2022; Buening et al., 2023; Shen & Huan, 2023; Cheng & Huan, 2025; Shen et al., 2025). Among these, the closest works to ours (Shen & Huan, 2023; Cheng & Huan, 2025; Shen et al., 2025) model a Bayesian agent's experiment selection problem using (partially observable) Markov decision processes. However, their objective is to design policies that maximize information gain while trading off experimental costs. In contrast, in the approval setting we study, the agent has direct economic incentives to conduct experiments, as product approval is financially beneficial.

2. A Protocol for Subsidized Sequential Experimentation

We consider an agent (the product developer) who seeks regulatory approval for a product from a principal (the regulator). The product is characterized by an efficacy parameter $\theta^* \in [0, 1]$, unknown both to the agent and the principal, with higher values indicating a more effective product.[1] Motivated by multi-stage clinical trials (U.S. Food and Drug Administration, 2019; Janiaud et al., 2021), in this section, we introduce a sequential approval process in which the agent conducts a sequence of RCTs to provide sufficient evidence that the product meets the principal's standards. In turn, the principal commits to subsidize a fraction of the agent's total experimentation cost and, after each trial, they decide whether to approve the product or require the agent to conduct further experimentation.

The agent begins the (sequential) approval process with a prior belief B_0 about the efficacy θ^* of its product, which we model using a Beta distribution $B_0 = \mathrm{Beta}(\alpha_0, \beta_0)$, where $\alpha_0, \beta_0 > 0$ are given parameters.[2] At each time step $t \in [T] = \{0, 1, \ldots, T\}$ of the approval process, the agent's action is to either (i) conduct a randomized trial with a sample size $n_t \in \{1, \ldots, n^{\mathtt{max}}\}$ and incur a cost $c(n_t)$, where $c : \mathbb{N} \to \mathbb{R}_+$ is a non-decreasing cost function, or (ii) opt out and stop the approval process at no additional cost, which we represent as $n_t = 0$ with $c(0) = 0$. If the agent decides to conduct a trial by selecting $n_t > 0$, the agent observes a random outcome $X_t \sim \mathrm{Bin}(n_t, \theta^*)$, which represents the number of *successes* in the trial and depends on the unknown efficacy θ^*, and this outcome is

[1] In certain settings, the principal may be interested not only in the efficacy of a product but also in other properties of the product such as its safety. In Appendix C, we discuss how to extend the approval process to such settings.

[2] The parameters α_0 and β_0 characterize prior information the agent may have about the product based on, *e.g.*, preliminary tests. In Appendix C, we show that the approval process can be extended to more general settings with arbitrary prior beliefs.

then revealed to the principal.[3] Then, based on the outcome X_t, the agent updates their (posterior) belief B_t about the true efficacy θ^* of the product, *i.e.*,

$$B_{t+1} = \mathrm{Beta}(\underbrace{\alpha_t + X_t}_{\alpha_{t+1}}, \underbrace{\beta_t + n_t - X_t}_{\beta_{t+1}}). \tag{1}$$

Throughout the process, the agent employs a (possibly randomized) policy $\pi \in \Pi$ to select the sample sizes $n_t \sim \pi(\alpha_t, \beta_t, C_t, t)$ based on their belief $B_t = \mathrm{Beta}(\alpha_t, \beta_t)$ about the efficacy of the product, the total cost $C_t = \sum_{k=0}^{t-1} c(n_k)$ they have incurred so far, and the time step t of the approval process. Moreover, the principal uses the revealed outcomes X_t as evidence to decide on the approval of the product. Formally, the principal conducts a (sequential) hypothesis test with null and alternative hypotheses given by

$$H_0 = \left\{\theta^* : \theta^* < \theta^{\flat}\right\} \text{ and } H_1 = \left\{\theta^* : \theta^* \geq \theta^{\flat}\right\}, \tag{2}$$

where $\theta^{\flat} \in (0, 1)$ is a baseline efficacy mandated by the principal and known to the agent (*e.g.*, the efficacy of the current standard-of-care treatment in the context of clinical trials). The null hypothesis H_0 therefore corresponds to the product failing to meet the principal's standard, and we will equivalently refer to the principal approving the product as *rejecting* H_0.

In general, the principal can implement any decision rule to reject H_0. However, in what follows, we draw on the literature on sequential hypothesis testing (Wald, 1945; Ramdas et al., 2023; Ramdas & Wang, 2025), and consider a principal who aims to control the false positive rate, *i.e.*, the probability of approving a product whose efficacy does not exceed the baseline $\theta^{\flat}$. Concretely, based on the sample size n_t and revealed outcome X_t, the principal computes—and shares with the agent—a non-negative quantity $E(X_t, n_t) \in \mathbb{R}_+$, referred to as an *e-value*, quantifying the observed evidence against H_0 at time t (*i.e.*, larger e-values correspond to stronger evidence against H_0).[4] Then, the principal aggregates all available evidence multiplicatively via a (stochastic) process M, which we will refer to as the *test process*, and whose value at time t is defined as:

$$M_t = \begin{cases} 1 & t = 0 \\ E(X_{t-1}, n_{t-1}) \cdot M_{t-1} & t \geq 1. \end{cases} \tag{3}$$

Finally, based on the test process M, the principal rejects the null hypothesis H_0 (approves the product) as soon as the accumulated evidence exceeds a fixed threshold $\kappa \in (0, 1)$ set in advance, *i.e.*, as soon as $M_{t+1} \geq 1/\kappa$ for some time step t. Here, κ acts as a tolerance parameter specifying how much evidence must be accumulated before rejecting H_0, where smaller values of κ correspond to a more conservative approval standard set by the principal, and note that, if the process fails to yield sufficient evidence for approval within T time steps, it concludes without approval.

In the above sequential test, the false positive rate is bounded by κ as long as the e-values are such that $\mathbb{E}_{H_0}[E(X_t, n_t)] \leq 1$, as shown elsewhere (Ramdas et al., 2023), *i.e.*,

$$P_{H_0}\left(\{\exists t \in [T] \colon M_{t+1} \geq 1/\kappa\}\right) \leq \kappa. \tag{4}$$

Importantly, the above guarantee on the false positive rate holds throughout the entire approval process—a particularly desirable property known as *any-time validity* (Grünwald et al., 2024). To ensure that $\mathbb{E}_{H_0}[E(X_t, n_t)] \leq 1$ holds, we construct an e-value that exponentiates a sufficient statistic for the unknown efficacy θ^* (de la Peña, 1999; Howard et al., 2020) by comparing the empirical success rate X_t/n_t against a monotone transformation of the baseline $\theta^{\flat}$:[5]

Proposition 1. *Given the (unknown) efficacy parameter $\theta^* \in [0, 1]$, $n_t > 0$, and a binomial variable $X_t \sim \mathrm{Bin}(n_t, \theta^*)$, the positive random variable*

$$E(X_t, n_t) = \exp\left(X_t - n_t \cdot \log(1 + \theta^{\flat}(e - 1))\right) \tag{5}$$

is a valid e-value under $H_0 = \left\{\theta^* : \theta^* < \theta^{\flat}\right\}$, that is, $\mathbb{E}_{X_t \sim \mathrm{Bin}(n_t, \theta^*)}[E(X_t, n_t)] \leq 1$ for any $\theta^* \in H_0$.

Furthermore, given the specific form of e-value in Eq. 5, it is easy to verify that the value of the test process M_t is uniquely determined by the parameters α_t and β_t characterizing the agent's belief:

Proposition 2. *At any time step $t \in [T]$, the value of the test process M_t satisfies:*

$$\begin{aligned} M_t &= f(\alpha_t, \beta_t) \quad \textit{where} \\ f(\alpha, \beta) &= \exp\big(\alpha - \alpha_0 \\ &\quad - (\alpha + \beta - \alpha_0 - \beta_0) \log(1 + \theta^{\flat}(e - 1))\big). \end{aligned} \tag{6}$$

If the agent's product is approved, the agent and the principal obtain benefits $\rho^{\mathrm{A}}, \rho^{\mathrm{S}} > 0$, respectively, and we refer to the latter as the social benefit upon approval. In the context of clinical trials, ρ^{A} can be interpreted as the (estimated) economic benefit obtained by the agent—the pharmaceutical company—from drug sales if the drug is approved, while ρ^{S} represents the corresponding benefit to society resulting from bringing an effective treatment to market.

Since the principal also benefits from a product's approval, it can be in their interest to incentivize the agent to continue

[3] The specific meaning of the number of successes is application dependent. In the context of clinical trials, it may correspond to the number of patients who recover after receiving a treatment.

[4] Computing $E(X_t, n_t)$ plays a role similar to that of a p-value for rejecting H_0, while offering stronger guarantees in sequential settings. See Appendix B for a brief overview of sequential hypothesis testing with e-values.

[5] See Appendix C for an extension to arbitrary e-values; proofs are deferred to Appendix F.

experimenting, particularly in situations where the agent's benefit $\rho^{\mathtt{A}}$ is not high enough to compensate for the total experimentation cost required to reject H_0. In the following, we focus our attention on a natural and widely-used mechanism through which the principal can provide such an incentive: subsidizing, conditional on approval, a fraction $\varepsilon \in [0, \varepsilon^{\mathtt{max}}]$ of the total cost incurred by the agent, where $\varepsilon^{\mathtt{max}} \leq 1$ denotes the maximum fraction the principal is willing to subsidize (Renwick et al., 2016; Wu & Zhao, 2022).

Given the above benefits, experimentation cost, and subsidies, the agent's and the principal's (expected) utilities $U^{\mathtt{A}}$ and $U^{\mathtt{S}}$ are given by

$$U^{\mathtt{A}}(\pi;\varepsilon) = \mathbb{E}_\pi\Bigg[\left(\rho^{\mathtt{A}} + \varepsilon\sum_{t=0}^{\tau} c(n_t)\right)\mathbb{1}\{M_{\tau+1} \geq 1/\kappa\} - \sum_{t=0}^{\tau} c(n_t)\,\Bigg|\,\theta^*\Bigg],$$
$$U^{\mathtt{S}}(\varepsilon;\pi) = \mathbb{E}_\pi\Bigg[\left(\rho^{\mathtt{S}} - \varepsilon\sum_{t=0}^{\tau} c(n_t)\right)\mathbb{1}\{M_{\tau+1} \geq 1/\kappa\}\,\Bigg|\,\theta^*\Bigg]. \tag{7}$$

where $\tau = T \wedge \min\{t \in [T] \colon n_t = 0 \text{ or } M_{t+1} \geq 1/\kappa\}$ is the last step of the approval process, $\mathbb{1}\{\bullet\}$ is the indicator function, and the expectation $\mathbb{E}_\pi[\bullet|\theta^*]$ is taken over the random outcomes $X_t \sim \mathrm{Bin}(n_t, \theta^*)$.[6]

However, since the agent and the principal do not know the true efficacy θ^* a priori, they cannot find the policy π and subsidy ε that maximize their respective utilities, as defined in Eq. 7. In the next sections, we investigate how the agent and principal may leverage their evolving beliefs about efficacy to find near-optimal policies and subsidies as the approval process progresses.

3. Experimental Design Using Belief Markov Decision Processes

If the agent does not know the true efficacy θ^*, we argue that, as the approval process progresses, they act based on their beliefs regarding θ^* given the trial outcomes observed (Harsanyi, 1979). Consequently, we consider an agent who determines their trial sample sizes by planning ahead—not according to the (unknown) actual evolution of the process, but according to how they anticipate their beliefs will evolve depending on their actions and anticipated outcomes.

To formalize the agent's planning strategy, we employ the framework of (belief) Markov decision processes (Sutton & Barto, 2018; Kaelbling et al., 1998). Specifically, for a fixed subsidy ε specified by the principal, we define the process $\mathcal{M}^\varepsilon = (\mathcal{S}, \mathcal{A}, P, r^\varepsilon, T)$, whose components we describe next.

The state space $\mathcal{S} = \mathcal{S}^{\mathtt{in}} \cup \{S^{\mathtt{out}}\}$ includes states $S = (\alpha, \beta, C) \in \mathcal{S}^{\mathtt{in}}$, where α and β are the parameters characterizing the agent's belief and C is the agent's total running cost, as well as a special absorbing state $S^{\mathtt{out}}$, which indicates that the agent has opted out of the approval process. The action space $\mathcal{A} = \{0, \ldots, n^{\mathtt{max}}\}$ consists of all possible sample sizes the agent may select.

Further, the transition distribution P characterizes how the agent anticipates their belief and cost will evolve after they conduct a trial of size n. Formally, a transition from a state $S = (\alpha, \beta, C)$ to a state S' follows from the (randomized) assignment

$$S' = \begin{cases} (\alpha + X,\ \beta + n - X,\ C + c(n)) & \text{if } n > 0,\ f(S) < 1/\kappa, \\ S & \text{if } n > 0,\ f(S) \geq 1/\kappa, \\ S^{\mathtt{out}} & \text{if } n = 0. \end{cases} \tag{8}$$

where $X \sim \mathrm{Bin}(n, \theta)$ with $\theta \sim \mathrm{Beta}(\alpha, \beta)$, and the function $f(S) = f(\alpha, \beta)$ as in Eq. 6 if $S \in \mathcal{S}^{\mathtt{in}}$ and $f(S^{\mathtt{out}}) = 0$. In words, the upper case captures the Bayesian belief update and the additional cost incurred by the agent for conducting an experiment of size n and observing an outcome X (see Eq. 1), the middle case captures the successful conclusion of the approval process, and the lower case captures the scenario in which the agent decides to opt out. Importantly, note that evaluating the likelihood of such transitions does not depend on the efficacy θ^* and thus is possible solely based on the agent's belief.

Finally, the reward r^ε characterizes the agent's anticipated profit or loss due to conducting a trial under subsidy ε. Formally, the reward for transitioning from state S to state S' via action n is given by

$$r^\varepsilon(S, n, S') = \Big[-c(n) + \big(\rho^{\mathtt{A}} + \varepsilon(C + c(n))\big) \mathbb{1}\{f(S') \geq 1/\kappa\}\Big] \cdot \mathbb{1}\{S \neq S^{\mathtt{out}},\ f(S) < 1/\kappa\}. \tag{9}$$

In words, as long as the approval process is in progress, the reward is simply equal to the (negative) cost $c(n)$ the agent anticipates to incur by conducting a trial with their chosen sample size n. If the updated state S' leads to product approval (*i.e.*, $f(S) < 1/\kappa$ and $f(S') \geq 1/\kappa$), the agent anticipates receiving their one-off benefit $\rho^{\mathtt{A}} > 0$ and a subsidized fraction of their total incurred cost throughout the process, as determined by the subsidy ε selected by the principal.

Now that we have defined all the components of the Markov decision process $\mathcal{M}^\varepsilon$, we can formalize the agent's strategy

[6] Throughout, we adopt the convention $\min \emptyset = +\infty$ and write $x \wedge y = \min(x, y)$ for $x, y \in \mathbb{R} \cup \{+\infty\}$.

within the process. Starting from the initial state $S_0 = (\alpha_0, \beta_0, 0)$, the agent aims to select a policy $\pi : \mathcal{S} \times [T] \to \Delta(\mathcal{A})$ maximizing their *anticipated utility*

$$\bar{U}^{\mathrm{A}}(\pi; \varepsilon) = \mathbb{E}_\pi \left[\sum_{t=0}^{T} r^\varepsilon(S_t, n_t, S_{t+1}) \middle| S_0 = (\alpha_0, \beta_0, 0) \right], \tag{10}$$

where the expectation is taken over the state transitions that the agent anticipates to occur throughout the approval process, given their initial belief. Here, it is important to note that both the reward r^ε and the transition distribution P are known to the agent, hence they can evaluate the anticipated utility of any given policy π before the approval process starts. Moreover, due to Bellman's optimality principle in MDPs (Sutton & Barto, 2018), for any $S \in \mathcal{S}$ and time step $l \in [T]$, the agent's optimal policy $\pi^\varepsilon \in \operatorname{argmax}_\pi \bar{U}^{\mathrm{A}}(\pi; \varepsilon)$ satisfies that

$$\pi^\varepsilon(S, l) \in \operatorname*{argmax}_\pi V_\pi^\varepsilon(S, l)$$
$$\text{where} \quad V_\pi^\varepsilon(S, l) = \mathbb{E}_\pi \left[\sum_{t=l}^{T} r^\varepsilon(S_t, n_t, S_{t+1}) \middle| S_l = S \right]. \tag{11}$$

In the above equation, the function V_π^ε is often referred as the value function. Moreover, note that $\bar{U}^{\mathrm{A}}(\pi; \varepsilon) = V_\pi^\varepsilon(\alpha_0, \beta_0, 0, 0)$.

In this context, a natural question is how the anticipated utility $\bar{U}^{\mathrm{A}}(\pi; \varepsilon)$ relates to the agent's true utility $U^{\mathrm{A}}(\pi; \varepsilon)$ defined in Eq. 7. The following proposition shows that the anticipated utility is equal to true agent utility averaged over the agent's initial belief about the true efficacy θ^* of their product at the start of the approval process.

Proposition 3. *Let $\varepsilon \in [0, \varepsilon^{max}]$ be any subsidy set by the principal, and π be any agent policy. Then,*

$$\bar{U}^{\mathrm{A}}(\pi; \varepsilon) = \mathbb{E}_{\theta^* \sim B_0} \left[U^{\mathrm{A}}(\pi; \varepsilon) \right]. \tag{12}$$

The above proposition reveals that an agent implementing the decision policy π^ε is, in fact, a Bayesian decision maker maximizing their expected utility (Ghavamzadeh et al., 2015; Russo et al., 2020; Buening et al., 2023).

Next, we formalize the principal's strategy within the process. Let Q be a prior distribution characterizing the principal's prior knowledge about the agent's initial belief (α_0, β_0). Then, the principal aims to select a subsidy ε maximizing the *anticipated social utility*

$$\bar{U}^{\mathrm{S}}(\varepsilon; \pi) = \mathbb{E}_{(\alpha_0, \beta_0) \sim Q} \Bigg[\mathbb{E}_\pi \Bigg[\sum_{t=0}^{T} \left(\rho^{\mathrm{S}} - \varepsilon \cdot C_{t+1} \right) \cdot \mathbb{1}\left\{ 0 < f(S_t) < 1/\kappa \leq f(S_{t+1}) \right\} \Bigg| S_0 \Bigg] \Bigg]. \tag{13}$$

where the inner expectation is taken over the state transitions that the principal anticipates to occur throughout the approval process, given the prior knowledge about the agent's initial belief.

In the next sections, we analyze how the agent and the principal maximize their anticipated utilities $\bar{U}^{\mathrm{A}}(\pi; \varepsilon)$ and $\bar{U}^{\mathrm{S}}(\varepsilon; \pi)$ within a natural Stackelberg setting (von Stackelberg, 2010). In this setting, the principal (the leader) first commits to a subsidy level ε^* maximizing the social utility $\bar{U}^{\mathrm{S}}$, accounting for the fact that, given any ε, the agent (the follower) will implement the optimal policy π^ε that maximizes their utility $\bar{U}^{\mathrm{A}}$. Then, after observing this commitment, the agent implements the optimal policy. Formally, the subsidy level ε^* is the solution to the following optimization problem:

$$\begin{aligned} \varepsilon^* &= \operatorname*{argmax}_{\varepsilon \in [0, \varepsilon^{\max}]} \bar{U}^{\mathrm{S}}\left(\varepsilon; \pi^\varepsilon\right) \\ &\text{subject to} \quad \pi^\varepsilon \in \operatorname*{argmax}_{\pi \in \Pi} \bar{U}^{\mathrm{A}}(\pi; \varepsilon). \end{aligned} \tag{14}$$

The above Stackelberg setting fits a variety of real-world applications. For example, in clinical trials, funding agencies act as leaders who commit to and announce a funding program; drug developers are the followers who apply to the program and, if successful, run a trial (U.S. Congress, 2024; National Institutes of Health, 2026b).

4. Optimal Experimental Design under Subsidies

In this section, we characterize the agent's optimal policy π^ε along with its corresponding value function

$$V^\varepsilon(S, l) = \max_{\pi \in \Pi} V_\pi^\varepsilon(S, l) \quad \forall\, S \in \mathcal{S},\, l \in [T]. \tag{15}$$

To this end, we first note that, once the process $\mathcal{M}^\varepsilon$ reaches a state where $f(S) \geq 1/\kappa$, or the state $\mathcal{S}^{\mathtt{out}}$, the action $n = 0$ is always optimal. This is implied by the transition dynamics and the reward definition in Eqs. 8 and 9. Therefore, to find the policy π^ε, it suffices to consider states S where $0 < f(S) < 1/\kappa$. Further, we show that, in the process $\mathcal{M}^\varepsilon$, the number of (unique) reachable states S where $f(S) \geq 1/\kappa$ is finite:

Proposition 4. *For any realization of the process $\mathcal{M}^\varepsilon$, any reached state S where $0 < f(S) < 1/\kappa$ belongs to a finite set $S^r \subset \mathcal{S}$. Moreover, if the cost function c is linear, the set S^r has size $\mathcal{O}((n^{max})^2 \cdot T^3)$.*

As an immediate consequence, we can find an optimal (deterministic) policy π^ε using standard planning methods. In particular, in Algorithm 2, we provide an adaptation of the classical value iteration algorithm (Hernández-Lerma & Lasserre, 1996), which is guaranteed to find an optimal policy π^ε in $\mathcal{O}((n^{\max})^4 \cdot T^3)$. In the remainder of this section,

we derive several key structural insights about the optimal policy π^ε and the optimal value function V^ε, which will be helpful to efficiently find the principal's optimal subsidy.

Our starting point is the observation that, in light of Proposition 2, the value of the test process M at a state $S = (\alpha, \beta, C)$ increases with α and decreases with β. This suggests that states with larger α and smaller β are more favorable. The following proposition formalizes this intuition by showing that the optimal value function V^ε satisfies a monotonicity property in both the belief parameters and the accumulated cost.

Proposition 5. *For any time step $l \in [T]$ and for any pair of states $S = (\alpha, \beta, C) \in \mathcal{S}^{in}$ and $S' = (\alpha', \beta', C') \in \mathcal{S}^{in}$ such that $f(S) < 1/\kappa$ and $f(S') < 1/\kappa$, the following holds:*

1. $V^\varepsilon(\alpha, \beta, C, l) \leq V^\varepsilon(\alpha', \beta, C, l)$ *if* $\alpha \leq \alpha'$;
2. $V^\varepsilon(\alpha, \beta, C, l) \geq V^\varepsilon(\alpha, \beta', C, l)$ *if* $\beta \leq \beta'$;
3. $V^\varepsilon(\alpha, \beta, C, l) \leq V^\varepsilon(\alpha, \beta, C', l)$ *if* $C \leq C'$.

Leveraging the above proposition, we can characterize the conditions under which the optimal policy π^ε opts out of the approval process by selecting the action $n = 0$. In particular, the following proposition shows that π^ε exhibits a threshold structure: at a state with belief (α, β), it selects $n = 0$ if and only if β exceeds a threshold that depends on α (see Figure 3 in Appendix D for an illustration).

Proposition 6. *For any time step $t \in [T]$ and total cost C_t, there exists a non-decreasing function $\tilde{\beta} : \mathbb{R}_+ \to \mathbb{R}_+$ such that, for any state $S_t = (\alpha_t, \beta_t, C_t) \in \mathcal{S}^{in}$, the optimal policy π^ε opts out of the approval process if $\beta_t > \tilde{\beta}(\alpha_t)$, and it does not opt out if $\beta_t < \tilde{\beta}(\alpha_t)$.*

Furthermore, we can also characterize how the agent's anticipated utility $\bar{U}^{\mathrm{A}}(\pi^\varepsilon; \varepsilon)$ under the optimal policy π^ε depends on the subsidy. To this end, we first show that, for any policy π, the value function V^ε_π is linear in the subsidy ε:

Proposition 7. *For any policy π, state $S \in \mathcal{S}$, and time step $l \in [T]$, the value function V^ε_π admits a linear decomposition*

$$V^\varepsilon_\pi(S, l) = V^0_\pi(S, l) + \varepsilon \cdot A_\pi(S, l), \tag{16}$$

where $V^0_\pi(S, l)$ is the value function in the unsubsidized process $\mathcal{M}^0$, and $A_\pi(S, l) \geq 0$ denotes the expected total cost incurred conditional on approval, starting from state S at time l (see Eq. 56 in Appendix F.9).

Building upon this result, the agent's anticipated utility $\bar{U}^{\mathrm{A}}(\pi^\varepsilon; \varepsilon)$ under the optimal policy π^ε admits a concise structural characterization, as formalized by the following proposition:

Proposition 8. *The agent's optimal anticipated utility $\bar{U}^{\mathrm{A}}(\pi^\varepsilon; \varepsilon)$, and its expectation $\mathbb{E}_{(\alpha_0, \beta_0) \sim Q}[\bar{U}^{\mathrm{A}}(\pi^\varepsilon; \varepsilon)]$ according to the principal's belief Q, are piecewise linear, continuous, and convex functions of the subsidy ε over a partition $\mathcal{P} = \{\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_L\}$ of the interval $[0, \varepsilon^{\max}]$, with $0 = \varepsilon_0 < \varepsilon_1 < \cdots < \varepsilon_L = \varepsilon^{\max}$. Moreover, for each interval of $\mathcal{P}$, the agent's optimal policy is constant, that is, $\pi^\varepsilon = \pi_i$ for all $\varepsilon \in [\varepsilon_i, \varepsilon_{i+1})$.*

In the next section, we leverage the above characterization of the agent's anticipated optimal utility to develop an algorithm that computes the principal's optimal subsidy ε^*, as defined in Eq. 14.[7]

5. Finding Optimal Subsidies

To find the optimal subsidy ε^*, our starting point is the observation that, for any policy π, the anticipated social utility $\bar{U}^{\mathrm{S}}(\varepsilon; \pi)$ is a decreasing function of ε. This is because, as the subsidy ε increases, the principal covers a higher fraction of the agent's cost, but the probability that the product receives approval remains unchanged. More formally, we have the following proposition:

Proposition 9. *For any policy π and subsidy $\varepsilon \in [0, \varepsilon^{max}]$, the anticipated social utility $\bar{U}^S(\varepsilon; \pi)$ admits a linear decomposition*

$$\bar{U}^S(\varepsilon; \pi) = \rho^S \mathbb{E}_{(\alpha_0, \beta_0) \sim Q}\big[P_\pi(\mathcal{E} \mid S_0)\big] - \varepsilon\, \mathbb{E}_{(\alpha_0, \beta_0) \sim Q}\big[A_\pi(S_0, 0)\big],$$

where $\mathcal{E} = \{\exists t \in [T] \colon f(S_{t+1}) \geq 1/\kappa\}$ denotes the approval event.

In the above expression, the first term corresponds to the principal's anticipated probability of rejecting H_0 under policy π, and the second term corresponds to the anticipated total cost borne by the principal through the subsidy.

As a consequence, and in light of Proposition 8, in each interval $[\varepsilon_i, \varepsilon_{i+1})$ of the partition $\mathcal{P}$ where a fixed policy π_i is optimal, the anticipated social utility $\bar{U}^{\mathrm{S}}(\varepsilon; \pi_i)$ is a decreasing (linear) function of the subsidy over $[\varepsilon_i, \varepsilon_{i+1})$, and therefore it is maximized at the left point ε_i:

$$\max_{\varepsilon \in [\varepsilon_i, \varepsilon_{i+1})} \bar{U}^{\mathrm{S}}(\varepsilon; \pi_i) = \bar{U}^{\mathrm{S}}(\varepsilon_i; \pi_i). \tag{17}$$

Crucially, the principal can compute each policy π_i *without knowing the agent's initial belief* (α_0, β_0), since the optimal policy for the process $\mathcal{M}^{\varepsilon_i}$ does not depend on the initial state. This stands in contrast to many settings in the

[7] In principle, Eq. 14 may allow for multiple solutions. However, our objective is not to characterize the full set of solutions, but rather to provide an algorithmic procedure to compute one such solution.

Algorithm 1 Finds the Principal's Optimal Subsidy

1: **Input:** MDP solver `SolveMDP`, maximum subsidy $\varepsilon^{\max}$, principal's belief Q
2: **Initialize:** $\mathcal{I} \leftarrow \emptyset$; $\varepsilon_L \leftarrow 0$; $\varepsilon_R \leftarrow \varepsilon^{\max}$
3: $(\pi_L, V_L^0, A_L) \leftarrow \texttt{SolveMDP}(\mathcal{M}^0)$
4: $(\pi_R, V_R^0, A_R) \leftarrow \texttt{SolveMDP}(\mathcal{M}^{\varepsilon^{\max}})$
5: $\bar{V}_j^0 \leftarrow \mathbb{E}_Q[V_j^0(\alpha_0, \beta_0, 0, 0)]$, $\bar{A}_j \leftarrow \mathbb{E}_Q[A_j(\alpha_0, \beta_0, 0, 0)]$ for $j \in \{L, R\}$
6: $\mathcal{U} \leftarrow \{(0, \bar{U}^{\mathrm{S}}(0; \pi_L))\}$
7: push $(\varepsilon_L, \pi_L, \bar{V}_L^0, \bar{A}_L, \varepsilon_R, \pi_R, \bar{V}_R^0, \bar{A}_R)$ into $\mathcal{I}$
8: **while** $\mathcal{I} \neq \emptyset$ **do**
9: Pop $(\varepsilon_L, \pi_L, \bar{V}_L^0, \bar{A}_L, \varepsilon_R, \pi_R, \bar{V}_R^0, \bar{A}_R)$ from $\mathcal{I}$
10: **if** $A_L \neq A_R$ **then**
11: $\varepsilon_{int} \leftarrow (\bar{V}_L^0 - \bar{V}_R^0)/(\bar{A}_R - \bar{A}_L)$
12: $(\pi_{int}, V_{int}^0, A_{int}) \leftarrow \texttt{SolveMDP}(\mathcal{M}^{\varepsilon_{int}})$
13: $\bar{V}_{int}^0 \leftarrow \mathbb{E}_Q[V_{int}^0(\alpha_0, \beta_0, 0, 0)]$
14: $\bar{A}_{int} \leftarrow \mathbb{E}_Q[A_{int}(\alpha_0, \beta_0, 0, 0)]$
15: **if** $\bar{V}_{int}^0 + \varepsilon_{int} \cdot \bar{A}_{int} \leq \bar{V}_L^0 + \varepsilon_{int} \cdot \bar{A}_L$ **then**
16: $\mathcal{U} \leftarrow \mathcal{U} \cup \{(\varepsilon_{int}, \bar{U}^{\mathrm{S}}(\varepsilon_{int}; \pi_R))\}$
17: **else**
18: Push $(\varepsilon_L, \pi_L, \bar{V}_L^0, \bar{A}_L, \varepsilon_{int}, \pi_{int}, \bar{V}_{int}^0, \bar{A}_{int})$ and $(\varepsilon_{int}, \pi_{int}, \bar{V}_{int}^0, \bar{A}_{int}, \varepsilon_R, \pi_R, \bar{V}_R^0, \bar{A}_R)$ into $\mathcal{I}$
19: **end if**
20: **end if**
21: **end while**
22: **Return** ε^* where $(\varepsilon^*, u^*) = \operatorname{argmax}_{(\varepsilon,u)\in\mathcal{U}} u$

literature on (Bayesian) Stackelberg games, where the principal typically must anticipate the agent's best response by averaging over the agent's private information (Conitzer & Sandholm, 2006; Paruchuri et al., 2008).

Leveraging the above results, we derive an efficient divide-and-conquer procedure to find the optimal subsidy ε^*. The procedure (i) explicitly constructs the partition $\mathcal{P}$ and (ii) applies Proposition 9 to determine the optimal anticipated social utility on each interval $[\varepsilon_i, \varepsilon_{i+1})$ by evaluating it at the left endpoint ε_i. Algorithm 1 summarizes the overall procedure, and the following proposition establishes its correctness.

Proposition 10. *Algorithm 1 is guaranteed to find an optimal subsidy ε^* in a finite number of iterations.*

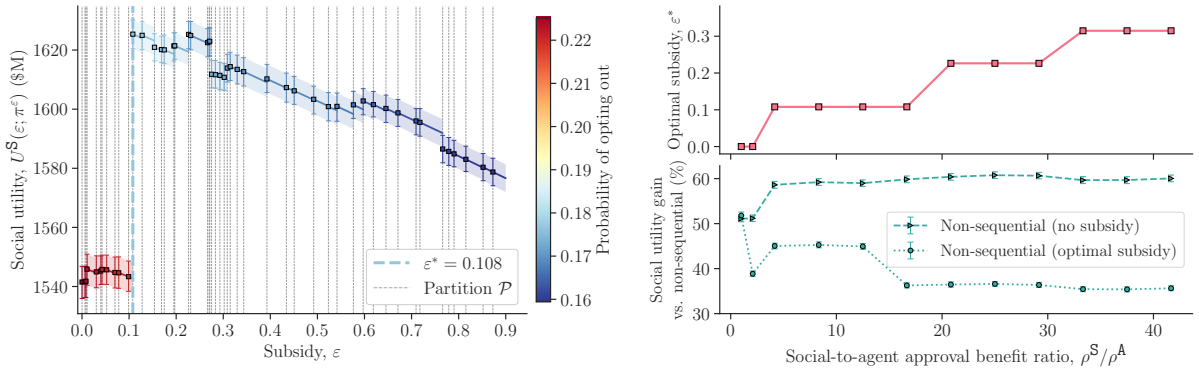
Algorithm 1 maintains a stack $\mathcal{I}$ of intervals, alongside their corresponding optimal policies at the endpoints and the linear value function decompositions provided by Proposition 7. We abstract the computation of these policies and value functions into a procedure, `SolveMDP`, which may be implemented using a value-iteration algorithm (e.g., Algorithm 3). The stack $\mathcal{I}$ is initialized to the full interval $[0, \varepsilon^{\max}]$ (lines 3 to 5). At each iteration, the algorithm calculates the intersection point ε_{int} between the endpoint value functions, $\bar{V}_L^0 + \varepsilon \cdot \bar{A}_L$ and $\bar{V}_R^0 + \varepsilon \cdot \bar{A}_R$ (line 9), where the bar denotes evaluation at S_0 and averaging over the principal's belief Q. Then, it computes the optimal value function at ε_{int} (line 10). Due to the convexity of the optimal value function, if the optimal value function at ε_{int} coincides with the value $\bar{V}_L^0 + \varepsilon_{int} \cdot \bar{A}_L = \bar{V}_R^0 + \varepsilon_{int} \cdot \bar{A}_R$, the policy π_L is optimal in $[\varepsilon_L, \varepsilon_{int})$, and the policy π_R is optimal in $[\varepsilon_{int}, \pi_R]$. In this case, ε_{int} is a vertex of $\mathcal{P}$, and the algorithm stores the social utility at ε_{int} (line 13). Conversely, if the optimal policy at ε_{int} strictly improves over the policy π_L, then the algorithm has found a new interval of the partition, and the stack $\mathcal{I}$ is updated (line 15). The algorithm then iterates the same steps over all intervals in the stack $\mathcal{I}$.[8]

6. Experiments: Subsidizing Antibiotic Development

Antimicrobial resistance is a major global threat, projected to cause 10 million deaths annually by 2050 (GBD 2021 Antimicrobial Resistance Collaborators, 2024; Shallcross et al., 2015). Yet, FDA antibiotic approvals have dropped from 13% of all drugs in 1980 to 4% in the 2000s (Outterson et al., 2013), largely for economic reasons: treatments are short, prices must stay low to ensure availability, use is restricted to limit resistance, and competition from existing or generic drugs is intense (Shlaes, 2019; Gargate et al., 2025). Consequently, many major pharmaceutical companies have exited or reduced antibiotic pipelines (Plackett, 2020), while small biotech firms often struggle financially (Courtemanche et al., 2021; Piddock et al., 2024; Wells et al., 2024). To counter this, public and private efforts have focused on incentivizing and subsidizing development (Outterson et al., 2016; Anderson et al., 2023; United States Congress, 2026). In this section, we conduct a series of experiments to demonstrate the effectiveness of our approval protocol in optimally subsidizing antibiotic development.

Experimental setup. The principal subsidizes the development of an antibiotic and conducts a hypothesis test with $\kappa = 0.05$ (*i.e.*, a false positive rate of at most 0.05) to determine whether the antibiotic's (unknown) efficacy $\theta^* = 0.65$ exceeds the benchmark $\theta^{\mathrm{b}} = 0.5$. The agent can conduct up to four trials ($T = 3$), each with a maximum sample size of $n^{\max} = 200$ patients. Although data on the economic cost and sales of antibiotic development are mostly private, recent reports estimate that the average present value of sales is approximately \$240 M (Rahman et al., 2021), the average per-patient Phase III cost is \$66 k (Stergiopoulos et al., 2018), and the average fixed cost per trial is \$48.9 M (Moore et al., 2018). Therefore, we set $\rho^{\mathbb{A}} = \$240\,\mathrm{M}$, and $c(n) = \$48.9\,\mathrm{M} + \$0.066\,\mathrm{M} \cdot n$ for any $n \in \{1, \ldots, n^{\max}\}$. Further, we assume the agent has a non-informative (uniform) prior with $\alpha_0 = 1$, $\beta_0 = 1$, known to the principal, and vary ρ^{S}. Refer to Appendix G for additional details regarding our experimental setup, and to

[8] We report runtime measurements of Algorithm 1 in Appendix G.



(a) Social utility vs. subsidy ($\rho^{\mathrm{S}} = \$2000\,\mathrm{M}$)

(b) Optimal subsidy and social utility gain vs. $\rho^{\mathrm{S}}/\rho^{\mathrm{A}}$

Figure 1. **Subsidizing antibiotic development.** The figure shows the results of the approval process for an antibiotic with true (unknown) efficacy $\theta^* = 0.65$. Panel (a) shows the result of running Algorithm 1 to compute the optimal subsidy for the principal $\varepsilon^* = 0.108$ when the social benefit of approval is $\rho^{\mathrm{S}} = \$2000\,\mathrm{M}$. The dashed vertical lines correspond to the intervals of the partition $\mathcal{P}$ where the agent's optimal policy is constant (Proposition 8), and the colors indicate the probability that the agent opts out by selecting $n = 0$ during the approval process (before the drug is approved). Panel (b) shows, as a function of the social-to-agent approval benefit ratio, the optimal subsidy, together with the percentage increase in social utility of the sequential approval protocol relative to a non-sequential approval protocol in which the agent is restricted to a single trial with $n^{\max} = 800$. The error bars represent 95% bootstrapped confidence intervals.

Appendix H for results under alternative parameter choices and extensive sensitivity analyses.

Results. For an antibiotic with a social benefit upon approval of $\rho^{\mathrm{S}} = \$2000\,\mathrm{M}$ (a ratio $\rho^{\mathrm{S}}/\rho^{\mathrm{A}} \approx 8.3$), Panel (a) of Figure 1 shows (i) the social utility $U^{\mathrm{S}}(\varepsilon;\pi^\varepsilon)$, (ii) the partition $\mathcal{P}$, which consists of 49 intervals, and (iii) the agent's opt-out probability, over the entire range of subsidy levels. We find that, under the optimal subsidy $\varepsilon^* = 0.108$, the social utility increases by $\sim$5.5% and the agent's opt-out probability decreases by $\sim$22% compared to a scenario with no subsidies. Interestingly, we also find that the optimal subsidy ε^*—which maximizes the anticipated social utility $\bar{U}^{\mathrm{S}}(\varepsilon;\pi^\varepsilon)$—also maximizes the true social utility $U^{\mathrm{S}}(\varepsilon;\pi^\varepsilon)$. Further, Panel (b) of Figure 1 shows the optimal subsidy ε^* and the social utility gain compared to two non-sequential baselines for different values of the social-to-agent approval benefit ratio $\rho^{\mathrm{S}}/\rho^{\mathrm{A}}$. We find that the optimal subsidy increases with the social-to-agent approval benefit ratio, reflecting that higher societal utility strengthens the principal's incentive to subsidize experimentation. We also find that, compared to two non-sequential baselines restricting the agent to a single clinical trial (with a larger maximum sample size of $n^{\max} = 800$), our approval protocol yields substantial gains in social utility. Relative to a non-sequential protocol without subsidies, it increases social utility by approximately 50%–60%, depending on the ratio $\rho^{\mathrm{S}}/\rho^{\mathrm{A}}$; relative to one with optimal subsidies, it still achieves gains exceeding 35%.

7. Discussion and Limitations

In this section, we highlight several limitations of our work and discuss avenues for future research.

Methodology. In our work, we have considered a Beta–Binomial model, which is particularly natural in the context of RCTs. However, in other application domains, it may be desirable to consider more general models where the agent's beliefs, the experimental outcomes, and the e-values exhibit greater complexity—for instance, through dependence across trials or through the inclusion of safety characteristics. In Appendix C, we outline how to extend our framework to these more general settings; however, this extension introduces significant computational challenges. Furthermore, although our results do not require the principal to know the agent's initial belief (*i.e.*, its private information B_0), observing the agent's experimental actions over time may provide information about its prior, suggesting that the principal could, in principle, dynamically update its belief about the agent's prior. It would be interesting to incorporate such learning and elicitation into a sequential approval protocol in future work (Shi et al., 2025). In addition, it would also be valuable to extend our protocol to allow for uncertainty in the agent's benefit ρ^{A}, which in practice may vary across product developers and contexts. Finally, real-world regulatory settings may involve additional non-economic factors that are not explicitly captured in our model. For instance, clinical trial duration, patient follow-up requirements, and other operational constraints can affect both the feasibility and optimality of sequential experimentation protocols, and accounting for these factors remains an important direction for future work.

Implementation and Evaluation. We have conducted a case study applying our approval protocol to antibiotic development, a well-known setting characterized by underinvestment and market failure, which is particularly suited to studying the effects of subsidies (see Appendix H for further experimental results using different parameter values). However, it would be interesting to extend this analysis to other settings—such as orphan drugs or rare disease treatments—where trial costs and sample sizes may differ substantially from those in standard antibiotic development and may therefore lead to different optimal subsidies. We also empirically evaluated the computational cost of computing the agent's optimal decision policy and the optimal subsidy. In our experiments, both procedures were efficient and typically completed within a few minutes (see Appendix G for more details). However, as shown theoretically in Section 4, the complexity of computing the optimal agent policy scales polynomially with the number of actions and trials. Consequently, extending our methodology to more general settings—such as those described in Appendix C—might require developing approximate algorithms, as computing (Stackelberg) equilibria in general Bayesian games is known to be computationally intractable in many cases (Conitzer & Sandholm, 2006).

8. Conclusions

How can approval protocols be designed to incentivize experimentation? We address this question by introducing a sequential approval protocol in which the agent continuously refines its knowledge about the product and the principal subsidizes a fraction of the agent's experimental costs, with anytime-valid guarantees on the false positive rate. We show that the agent can efficiently compute the optimal experimental policy, and the principal can find the socially optimal subsidy when anticipating that the agent acts strategically. Using real-world data on antibiotic development, we demonstrate gains of up to 60% in social utility over non-sequential protocols without subsidies. More broadly, we hope our work provides insights for designing approval protocols that better align agents' incentives with social objectives.

Acknowledgements

Gomez-Rodriguez acknowledges support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 945719 and 101169607). Tsirtsis acknowledges supports from the Alexander von Humboldt Foundation in the framework of the Alexander von Humboldt Professorship (Humboldt Professor of Technology and Regulation awarded to Sandra Wachter) endowed by the Federal Ministry of Education and Research via the Hasso Plattner Institute.

Impact Statement

Regulatory agencies and product developers are progressively exploring more flexible and data-driven approval methodologies, including Bayesian approaches, to better balance safety and innovation (U.S. Food and Drug Administration, 2026). Our work contributes to this direction by providing a principled framework for designing subsidy mechanisms that improve social utility in approval processes, which may inform future policy discussions and be of interest to both public and private regulators and decision makers.

References

Anderson, M., Panteli, D., van Kessel, R., Ljungqvist, G., Colombo, F., and Mossialos, E. Challenges and opportunities for incentivising antibiotic research and development in Europe. *Lancet Reg. Health Eur.*, 33(100705):100705, October 2023.

Bates, S., Jordan, M. I., Sklar, M., and Soloff, J. A. Principal-agent hypothesis testing. *arXiv preprint arXiv:2205.06812*, 2022.

Brown, C. H., Ten Have, T. R., Jo, B., Dagne, G., Wyman, P. A., Muthén, B., and Gibbons, R. D. Adaptive designs for randomized trials in public health. *Annual review of public health*, 30(1):1–25, 2009.

Buening, T. K., Dimitrakakis, C., Eriksson, H., Grover, D., and Jorge, E. Minimax-bayes reinforcement learning, 2023. URL https://arxiv.org/abs/2302.10831.

Campbell, B., Balakrishnan, N., and Vidakovic, B. *Encyclopedia of statistical sciences*. Methods and Applications of Statistics. John Wiley & Sons, Nashville, TN, 2 edition, December 2005.

Cheng, C. and Huan, X. Optimal stopping for sequential Bayesian experimental design. *arXiv preprint arXiv:2509.21734*, 2025.

Cho, N., Lee, S., Shin, H.-S., and Tsourdos, A. Bayesian learning approach to model predictive control. *arXiv preprint arXiv:2203.02720*, 2022.

Chugg, B., Gauthier, E., Jordan, M. I., Ramdas, A., and Waudby-Smith, I. Post-hoc large-sample statistical inference. *arXiv preprint arXiv:2603.08002*, 2026.

Conitzer, V. and Sandholm, T. Computing the optimal strategy to commit to. In *Proceedings of the 7th ACM Conference on Electronic Commerce*, EC '06, pp. 82–90,

New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595932364. doi: 10.1145/1134707.1134717. URL https://doi.org/10.1145/1134707.1134717.

Courtemanche, G., Wadanamby, R., Kiran, A., Toro-Alzate, L. F., Diggle, M., Chakraborty, D., Blocker, A., and van Dongen, M. Looking for solutions to the pitfalls of developing novel antibacterials in an economically challenging system. *Microbiol. Res. (Pavia)*, 12(1):173–185, March 2021.

de la Peña, V. H. A General Class of Exponential Inequalities for Martingales and Ratios. *The Annals of Probability*, 27(1):537 – 564, 1999. doi: 10.1214/aop/1022677271. URL https://doi.org/10.1214/aop/1022677271.

Detsky, A. S. Are clinical trials a cost-effective investment? *Jama*, 262(13):1795–1800, 1989.

Deutsche Forschungsgemeinschaft. Clinical trials. https://www.dfg.de/en/research-funding/funding-opportunities/programmes/individual/clinical-trials, 2024. Accessed: 2026-03-31.

Dhillon, G. S., Gonzalez, J., Pandeva, T., and Curth, A. E-scores for (in)correctness assessment of generative model outputs. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026. URL https://openreview.net/forum?id=PCRCLYgiVK.

European and Developing Countries Clinical Trials Partnership. The European and developing countries clinical trials partnership. https://www.edctp.org/, 2025. Accessed: 2026-03-31.

Farina, A., Moro, F., Fasslrinner, F., Sedghi, A., Bromley, M., and Siepmann, T. Strength of clinical evidence leading to approval of novel cancer medicines in Europe: A systematic review and data synthesis. *Pharmacology Research & Perspectives*, 9(4):e00816, 2021.

Frantz, S. Why are clinical costs so high? *Nature Reviews Drug Discovery*, 2(11), 2003.

Gargate, N., Laws, M., and Rahman, K. M. Current economic and regulatory challenges in developing antibiotics for gram-negative bacteria. *NPJ Antimicrob. Resist.*, 3(1):50, June 2025.

Gauthier, E., Bach, F., and Jordan, M. I. Backward conformal prediction. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=23ichdd74N.

Gauthier, E., Bach, F., and Jordan, M. I. Betting on equilibrium: Monitoring strategic behavior in multi-agent systems, 2026. URL https://arxiv.org/abs/2601.05427.

GBD 2021 Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance 1990-2021: a systematic analysis with forecasts to 2050. *Lancet*, 404(10459):1199–1226, September 2024.

Ghavamzadeh, M., Mannor, S., Pineau, J., and Tamar, A. Bayesian reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 8(5–6):359–483, November 2015. ISSN 1935-8245. doi: 10.1561/2200000049. URL http://dx.doi.org/10.1561/2200000049.

Gieringer, D. H. The safety and efficacy of new drug approval. *Cato J.*, 5:177, 1985.

Grossman, S. J. and Hart, O. D. An analysis of the principal-agent problem. In *Foundations of insurance economics: Readings in economics and finance*, pp. 302–340. Springer, 1992.

Grünwald, P., de Heide, R., and Koolen, W. Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1091–1128, 03 2024. ISSN 1369-7412. doi: 10.1093/jrsssb/qkae011. URL https://doi.org/10.1093/jrsssb/qkae011.

Harsanyi, J. C. Bayesian decision theory, rule utilitarianism, and Arrow's impossibility theorem. *Theory and Decision*, 11(3):289–317, 1979.

Hernández-Lerma, O. and Lasserre, J. B. *Discrete-Time Markov Control Processes*. Springer New York, 1996. ISBN 9781461207290. doi: 10.1007/978-1-4612-0729-0. URL http://dx.doi.org/10.1007/978-1-4612-0729-0.

Hossain, S., Chen, Y., and Chen, Y. Strategic hypothesis testing. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17(none):257 – 317, 2020. doi: 10.1214/18-PS321. URL https://doi.org/10.1214/18-PS321.

Huang, B., Xu, E., Ramchandran, K., Jiao, J., and Jordan, M. I. Towards anytime-valid statistical watermarking. *arXiv preprint arXiv:2602.17608*, 2026.

Janiaud, P., Irony, T., Russek-Cohen, E., and Goodman, S. N. U.S. Food and Drug Administration reasoning in approval

decisions when efficacy evidence is borderline, 2013-2018. *Ann. Intern. Med.*, 174(11):1603–1611, November 2021.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Mahajan, R. and Gupta, K. Adaptive design clinical trials: Methodology, challenges and prospect. *Indian journal of pharmacology*, 42(4):201–207, 2010.

Martin, L., Hutchens, M., Hawkins, C., and Radnov, A. How much do clinical trials cost? *Nature Reviews Drug Discovery*, 16(6):381–382, 2017.

McClellan, A. Experimentation and approval mechanisms. *Econometrica*, 90(5):2215–2247, 2022. doi: https://doi.org/10.3982/ECTA17021. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA17021.

Moore, T. J., Zhang, H., Anderson, G., and Alexander, G. C. Estimated costs of pivotal trials for novel therapeutic agents approved by the US food and drug administration, 2015-2016. *JAMA Intern. Med.*, 178(11):1451–1457, November 2018.

National Institutes of Health. Clinical trial-specific funding opportunities. https://grants.nih.gov/policy-and-compliance/policy-topics/clinical-trials/specific-funding-opportunities, 2026a. Accessed: 2026-03-31.

National Institutes of Health. Small business funding. https://seed.nih.gov/small-business-funding, 2026b. Accessed: 2026-05-03.

Outterson, K., Powers, J. H., Seoane-Vazquez, E., Rodriguez-Monguio, R., and Kesselheim, A. S. Approval and withdrawal of new antibiotics and other antiinfectives in the U.S., 1980-2009. *J. Law Med. Ethics*, 41(3):688–696, 2013.

Outterson, K., Rex, J. H., Jinks, T., Jackson, P., Hallinan, J., Karp, S., Hung, D. T., Franceschi, F., Merkeley, T., Houchens, C., Dixon, D. M., Kurilla, M. G., Aurigemma, R., and Larsen, J. Accelerating global innovation to address antibacterial resistance: introducing CARB-X. *Nat. Rev. Drug Discov.*, 15(9):589–590, September 2016.

O'Connor, A. B. Building comparative efficacy and tolerability into the fda approval process. *Jama*, 303(10):979–980, 2010.

Paruchuri, P., Pearce, J. P., Marecki, J., Tambe, M., Ordonez, F., and Kraus, S. Playing games for security: An efficient exact algorithm for solving Bayesian Stackelberg games. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 2*, pp. 895–902, 2008.

Peskir, G. and Shiryaev, A. N. *Optimal stopping and free-boundary problems*. Lectures in Mathematics. ETH Zürich. Birkhauser Verlag AG, Basel, Switzerland, 2006 edition, August 2006.

Piddock, L. J. V., Alimi, Y., Anderson, J., de Felice, D., Moore, C. E., Røttingen, J.-A., Skinner, H., and Beyer, P. Advancing global antibiotic research, development and access. *Nat. Med.*, 30(9):2432–2443, September 2024.

Plackett, B. Why big pharma has abandoned antibiotics. *Nature*, 586(7830):S50–S52, October 2020.

Powell, W. B. and Ryzhov, I. O. *Optimal Learning*. Wiley Series in Probability and Statistics. Wiley-Blackwell, Hoboken, NJ, March 2012.

Rahman, S., Lindahl, O., Morel, C. M., and Hollis, A. Market concentration of new antibiotic sales. *J. Antibiot. (Tokyo)*, 74(6):421–423, June 2021.

Ramdas, A. and Wang, R. Hypothesis testing with e-values. *Foundations and Trends in Statistics*, 1(1-2):1–390, 07 2025. ISSN 2978-4212. doi: 10.1561/3600000002. URL https://doi.org/10.1561/3600000002.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. Game-theoretic statistics and safe anytime-valid inference, 2023. URL https://arxiv.org/abs/2210.01948.

Renwick, M. J., Brogan, D. M., and Mossialos, E. A systematic review and critical assessment of incentive strategies for discovery and development of novel antibiotics. *J. Antibiot. (Tokyo)*, 69(2):73–88, February 2016.

Robbins, H. Optimal stopping. *The American Mathematical Monthly*, 77(4):333–343, 1970. ISSN 00029890, 19300972. URL http://www.jstor.org/stable/2316139.

Rockafellar, R. *Convex Analysis*. Princeton landmarks in mathematics and physics. Princeton University Press, 1970. ISBN 9780691080697. URL https://books.google.de/books?id=OI4Ph2dXXhsC.

Russo, D., Roy, B. V., Kazerouni, A., Osband, I., and Wen, Z. A tutorial on Thompson sampling, 2020. URL https://arxiv.org/abs/1707.02038.

Shallcross, L. J., Howard, S. J., Fowler, T., and Davies, S. C. Tackling the threat of antimicrobial resistance: from policy to sustainable action. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 370(1670):20140082, June 2015.

Shekhar, S. and Ramdas, A. Nonparametric two-sample testing by betting. *IEEE Trans. Inf. Theor.*, 70(2): 1178–1203, February 2024. ISSN 0018-9448. doi: 10.1109/TIT.2023.3305867. URL https://doi.org/10.1109/TIT.2023.3305867.

Shen, W. and Huan, X. Bayesian sequential optimal experimental design for nonlinear models using policy gradient reinforcement learning. *Computer Methods in Applied Mechanics and Engineering*, 416:116304, 2023.

Shen, W., Dong, J., and Huan, X. Variational sequential optimal experimental design using reinforcement learning. *Comput. Methods Appl. Mech. Eng.*, 444(118068): 118068, September 2025.

Shi, F. C., Bates, S., and Wainwright, M. J. Sharp results for hypothesis testing with risk-sensitive agents. *arXiv preprint arXiv:2412.16452*, 2024.

Shi, F. C., Wainwright, M. J., and Bates, S. Instance-adaptive hypothesis tests with heterogeneous agents. *arXiv preprint arXiv:2510.21178*, 2025.

Shin, J., Ramdas, A., and Rinaldo, A. E-detectors: A nonparametric framework for sequential change detection. *The New England Journal of Statistics in Data Science*, 2(2):229–260, 2024. ISSN 2693-7166. doi: 10.51387/23-NEJSDS51.

Shlaes, D. M. The economic conundrum for antibacterial drugs. *Antimicrob. Agents Chemother.*, 64(1), December 2019.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.*, 22(11):1359–1366, November 2011.

Stergiopoulos, S., Calvert, S. B., Brown, C. A., Awatin, J., Tenaerts, P., Holland, T. L., DiMasi, J. A., and Getz, K. A. Cost drivers of a hospital-acquired bacterial pneumonia and ventilator-associated bacterial pneumonia phase 3 clinical trial. *Clin. Infect. Dis.*, 66(1):72–80, January 2018.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018. URL http://incompleteideas.net/book/the-book-2nd.html.

Tetenov, A. An economic theory of statistical testing. CeMMAP working papers 50/16, Institute for Fiscal Studies, Sep 2016. URL https://ideas.repec.org/p/azt/cemmap/50-16.html.

United States Congress. H.R. 7352: PASTEUR Act of 2026, 2026. URL https://www.congress.gov/bill/119th-congress/house-bill/7352. To amend the Public Health Service Act to establish a program to develop innovative antimicrobial drugs.

U.S. Congress. 26 U.S.C. 45C — clinical testing expenses for certain drugs for rare diseases or conditions. Internal Revenue Code, 2024. URL https://www.govinfo.gov/app/details/USCODE-2024-title26/USCODE-2024-title26-subtitleA-chap1-subchapA-partIV-subpartD-sec45C. Accessed: 2026-05-03.

U.S. Food and Drug Administration. Demonstrating substantial evidence of effectiveness for human drug and biological products: Guidance for industry. Draft guidance, U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), December 2019. URL https://www.fda.gov/drugs/guidance-compliance-regulatory-information/guidances-drugs.

U.S. Food and Drug Administration. Clinical trials grants program. https://www.fda.gov/industry/orphan-products-grants-program/clinical-trials-grants-program, 2024. Accessed: 2026-03-31.

U.S. Food and Drug Administration. Use of Bayesian methodology in clinical trials of drug and biological products. Draft guidance, Center for Biologics Evaluation and Research and Center for Drug Evaluation and Research, Food and Drug Administration, March 2026.

Velasco, A. A., Tsirtsis, S., and Rodriguez, M. G. Auditing pay-per-token in large language models. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026. URL https://openreview.net/forum?id=6lboj007YA.

Ville, J. *Étude Critique de la Notion de Collectif*. Collection des monographies des probabilités. Gauthier-Villars, 1939. URL https://books.google.de/books?id=ETY7AQAAIAAJ.

von Stackelberg, H. *Market Structure and Equilibrium*. Springer, Berlin, Germany, 2011 edition, August 2010.

Wald, A. Sequential Tests of Statistical Hypotheses. *The Annals of Mathematical Statistics*, 16(2):117 – 186, 1945. doi: 10.1214/aoms/1177731118. URL https://doi.org/10.1214/aoms/1177731118.

Waudby-Smith, I. and Ramdas, A. Estimating means of bounded random variables by betting, 2022. URL https://arxiv.org/abs/2010.09686.

Waudby-Smith, I., Sandoval, R., and Jordan, M. I. Universal log-optimality for general classes of e-processes and sequential hypothesis tests, 2025. URL https://arxiv.org/abs/2504.02818.

Wells, N., Nguyen, V.-K., and Harbarth, S. Novel insights from financial analysis of the failure to commercialise plazomicin: Implications for the antibiotic investment ecosystem. *Humanit. Soc. Sci. Commun.*, 11(1), July 2024.

Wu, W.-S. and Zhao, K. Government r&d subsidies and enterprise r&d activities: theory and evidence. *Economic Research-Ekonomska Istraživanja*, 35(1):391–408, 2022. doi: 10.1080/1331677X.2021.1893204. URL https://doi.org/10.1080/1331677X.2021.1893204.

Xu, Z. and Ramdas, A. Online multiple testing with e-values. In Dasgupta, S., Mandt, S., and Li, Y. (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3997–4005. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/xu24a.html.

A. Summary of Notation

In Table 1 we summarize the key symbols used in the main body of the paper.

Table 1. **Summary of notation.**

Symbol	Description
κ	Principal's false positive rate bound
θ^*	True (unknown) product efficacy
θ^{b}	Baseline efficacy
H_0, H_1	Null and alternative hypotheses
t	Time index
l	Initial time index (if not 0)
T	Maximum number of trials (time horizon of the belief MDP)
ε	Principal's subsidy level
$\varepsilon^{\mathtt{max}}$	Maximum allowable subsidy
ε^*	Principal's optimal subsidy
B_t	Agent's belief at time t
α_t, β_t	Parameters of the agent's belief at time t
Q	Principal's belief over the agent's prior
n_t	Agent's action (sample size) at time t
$n^{\mathtt{max}}$	Maximum sample size
π	Agent's policy
π^ε	Optimal policy under subsidy ε
V^ε	Optimal value function under subsidy ε
V_π^ε	Value function of policy π under subsidy ε
A_π	Expected total cost conditional on approval under policy π
c	Cost function
X_t	Experimental outcome at time t
E_t	e-value at time t
M_t	Principal's test process at time t
f	Mapping from beliefs to the test process
C_t	Cumulative experimental cost at time t
$\mathcal{M}^\varepsilon$	Belief MDP under subsidy ε
r^ε	Reward function in the belief MDP under subsidy ε
S_t	State of $\mathcal{M}^\varepsilon$ at time t
τ	Stopping time (last non-absorbing state of $\mathcal{M}^\varepsilon$)
$\mathcal{P}$	Partition associated with fixed policies
ρ^{A}	Agent's approval benefit
ρ^{S}	Social approval benefit
U^{A}	Agent's utility under θ^*
U^{S}	Social utility under θ^*
$\bar{U}^{\mathrm{A}}$	Agent's anticipated utility computed via the belief MDP
$\bar{U}^{\mathrm{S}}$	Principal's anticipated social utility computed via the belief MDP

B. Background on Sequential Hypothesis Testing Using e-values

In this section, we provide a brief overview of sequential hypothesis testing using e-values. For a detailed exposition, we refer the reader to Ramdas et al. (Ramdas & Wang, 2025).

A central object in classical hypothesis testing is the *p-value*, used to assess a null hypothesis H_0 against an alternative H_1. Despite their widespread use across scientific disciplines, p-values suffer from important limitations when used without appropriate precautions. A canonical example is *p-hacking* (also called *sampling to a foregone conclusion*): if one repeatedly collects data, computes a p-value, and checks whether it falls below a fixed threshold (*e.g.*, 0.05), rejection of H_0 is eventually guaranteed even when H_0 is true (Campbell et al., 2005; Simmons et al., 2011).

Alongside classical methods based on p-values, a rich line of work has developed methods based on a game-theoretic formulation of statistics that is better suited to sequential hypothesis testing (Ramdas et al., 2023) and mitigates issues such as p-hacking. The key objects in this framework are *e-values*, which are nonnegative random variables E that satisfy:

$$\mathbb{E}_{H_0}[E] \leq 1, \tag{18}$$

where the expectation is taken under any distribution in H_0. Typically, as with p-values, e-values are computed from some observed data X that is intended to provide evidence to reject H_0; in the sequential approval protocol considered in Section 2, we make this dependence explicit by writing $E(n, X)$, where n is the sample size selected by the agent and X the number of successes in the control trial. The intuition behind e-values is straightforward: since E has expectation at most 1 under H_0, it can take large values only with small probability. Hence, observing a large e-value can be interpreted as evidence against H_0—and the larger the e-value, the stronger the evidence. Consequently, one can construct a statistical test that rejects H_0 whenever the e-value takes high values, *i.e.*, whenever $E \geq 1/\kappa$ for a certain threshold $\kappa \in (0, 1)$. More precisely, by Markov's inequality, this threshold automatically controls the false positive rate:

$$P_{H_0}(E \geq 1/\kappa) \leq \kappa \cdot \mathbb{E}_{H_0}[E] \leq \kappa. \tag{19}$$

Proposition 11. *If E is an e-value for H_0,* i.e., $\mathbb{E}_{H_0}[E] \leq 1$*, then the non-sequential hypothesis test*

$$\phi = \mathbb{1}\{E \geq 1/\kappa\}$$

is a level-κ test for H_0 for any $\kappa \in (0, 1)$: its false positive rate is at most κ.

Informally, e-values are as general as p-values in the sense that they exist under essentially the same technical conditions and can be transformed into one another. Their main advantage, however, arises in sequential settings—*i.e.*, when data arrive over time (Shin et al., 2024; Shekhar & Ramdas, 2024; Xu & Ramdas, 2024; Velasco et al., 2026; Gauthier et al., 2026)—as in clinical trials, which is the primary reason our formalism in Section 2 and 3 is built on e-values. One way to understand this advantage is to note that the defining condition of e-values (Eq. 18) is preserved under a wide range of operations. For example, convex combinations of e-values and multiplications of (independent) e-values remain valid e-values. In contrast, analogous operations do not generally preserve the validity of p-values.

To formalize the use of e-values in a sequential setting, let $\mathcal{F} = (\mathcal{F}_t)_{t=0}^{\infty}$ be a filtration on a given sample space, where each $\mathcal{F}_t$ is a σ-algebra representing the information available at time t. For instance, $\mathcal{F}$ may be the filtration generated by the observations $X_0, \ldots, X_t$. Then, given a sequential data stream $X_0, X_1, \ldots$, the goal of sequential hypothesis testing is to maintain and update a running measure of evidence against H_0 as new observations arrive. In the e-value framework, this can be achieved by constructing a sequence $E_0, E_1, \ldots$ adapted to $\mathcal{F}$—so that each E_t depends only on data observed up to time t—and satisfying

$$\mathbb{E}_{H_0}[E_t \mid \mathcal{F}_{t-1}] \leq 1. \tag{20}$$

The above condition is the sequential analogue of the defining property in Eq. 18. For instance, if the observations X_t are independent, one may construct each E_t from X_t alone, in which case the e-values $E_0, \ldots, E_t$ are mutually independent. In Section 4, we adopt this construction, but we emphasize that more complex e-values can be defined if the experimental protocol requires it. For example, in Eq. 27 we consider a data-dependent construction in which new evidence is collected only if a prior experiment satisfies certain conditions.

By interpreting the quantity E_t as the new evidence against H_0 obtained at time t, a canonical way to construct a sequential test is to define a stochastic test process $M = (M_t)_{t \geq 0}$ as

$$M_t = \begin{cases} 1 & t = 0 \\ E_t \cdot M_{t-1} & t \geq 1, \end{cases} \tag{21}$$

which simply corresponds to aggregating the previous e-values multiplicatively.[9] By contrast, combining p-values in a sequential or dependent setting typically requires specialized corrections to maintain validity (Grünwald et al., 2024). Whenever the e-values satisfy Eq. 20, the above process M forms a (non-negative) supermartingale under H_0, *i.e.*,

$$\mathbb{E}_{H_0}[M_t \mid \mathcal{F}_{t-1}] \leq M_{t-1}. \tag{22}$$

Intuitively, this property means that if H_0 is true, the value of M_t does not, in expectation, increase over time. Conversely, sustained growth of the process M_t provides evidence against H_0. Thus, a natural sequential test to reject H_0 is to monitor whether $M_t \geq 1/\kappa$ for a chosen $\kappa \in (0, 1)$, and to reject H_0 as soon as this condition is met. Ville's inequality (Ville, 1939), a sequential extension of Markov's inequality, guarantees that this procedure controls the false positive rate uniformly over time:

Theorem 12 (Ville's inequality)**.** *If the process M is a non-negative supermartingale, then,*

$$P_{H_0}(\exists t \in \mathbb{N} : M_t \geq 1/\kappa) = P_{H_0}\left(\sup_{t \geq 0} M_t \geq 1/\kappa\right) \leq \kappa. \tag{23}$$

Moreover, validity is preserved if the test is stopped at any stopping time τ adapted to $\mathcal{F}$ (but possibly depending on the observed data):

$$P_{H_0}(M_\tau \geq 1/\kappa) \leq \kappa, \tag{24}$$

which is a property known as *any-time validity* and provides a principled solution to the problem of p-hacking.

Although the above results hold for any choice of e-values satisfying Eq. 20, a natural practical question is which specific form to adopt. We discuss this, and how different choices affect the approval process, in Appendix C.

[9] In Eq. 29 of Appendix C we discuss a different method to construct a test process M without multiplying e-values.

C. Extension to Arbitrary Belief Functions and e-values

The approval protocol introduced in Section 2 can be formulated in a more general setting (at the expense of losing tractability), as we now outline.

Generalization of the approval process. In a general setting, the agent may begin the approval process with an arbitrary initial belief $B_0 \in \Delta(\Theta)$, where Θ denotes a general parameter space for the unknown parameter θ^* characterizing the product. At each step of the approval process, the agent selects an action $n_t \in \mathcal{A}^{\texttt{gen}} \cup \{0\}$, where $\mathcal{A}^{\texttt{gen}}$ is a (potentially infinite) set describing the design of the next randomized controlled trial (*e.g.*, sample size, participant characteristics, etc.), and $n_t = 0$ again denotes the option to opt out. If $n_t \neq 0$, an outcome $X_t \sim P(X_t \mid n_t, \theta^*)$ is observed and the agent incurs a cost $c(n_t)$, where $P(\cdot \mid n_t, \theta^*)$ is a likelihood function characterizing the data-collection process specific to each experimental setting and $c: \mathcal{A}^{\texttt{gen}} \to \mathbb{R}_+$ is a cost function. The agent then updates its belief via the Bayesian posterior:

$$B_{t+1}(\theta) = \frac{B_t(\theta) \cdot P(X_t \mid n_t, \theta)}{\int_\Theta B_t(\theta) \cdot P(X_t \mid n_t, \theta)\, d\theta}.$$

To decide on approval, the principal partitions the parameter space as $\Theta = \Theta_0 \sqcup \Theta_1$ and defines the following null and alternative hypotheses:

$$\begin{cases} H_0 = \{\theta^* \in \Theta_0\} & \text{(null)} \\ H_1 = \{\theta^* \in \Theta_1\} & \text{(alternative)}. \end{cases} \tag{25}$$

This formulation allows for richer principal objectives. For instance, the principal may be concerned not only with a drug's efficacy but also with its safety, in which case H_0 may correspond to treatments that are either insufficiently effective or unsafe. To conduct the above hypothesis test sequentially, at each time step, the principal may compute any test process value M_t, subject only to two conditions: (i) the test process M must be predictable with respect to the filtration $\mathcal{F} = (\mathcal{F}_t)_{t=0}^{\infty}$ generated by the data and agent actions $X_0, n_0, X_1, n_1, \ldots$, and (ii) the test process M must be a supermartingale under H_0. The principal can then reject H_0 whenever $M_t \geq 1/\kappa$ at any time step, as described in Appendix B. A particular construction of the test process that generalizes Eq. 3 proceeds multiplicatively as

$$M_t = \begin{cases} 1 & t = 0, \\ E_{t-1}(X_{t-1}, n_{t-1}) \cdot M_{t-1} & t \geq 1, \end{cases} \tag{26}$$

where each E_{t-1} is an e-value adapted to $\mathcal{F}$ that may depend on all previously observed experimental outcomes and actions, *i.e.*, $X_0, n_0, X_1, n_1, \ldots, X_{t-1}, n_{t-1}$, and satisfies:

$$\mathbb{E}_{H_0}[E_t(n_t, X_t) \mid \mathcal{F}_{t-1}] \leq 1.$$

This allows modeling adaptive experimental designs in which future trials depend on past outcomes. Such settings arise naturally in multi-stage clinical trials (U.S. Food and Drug Administration, 2019), where progression to a subsequent phase may be contingent on earlier success. This can be represented by defining:

$$E_t(n_t, X_t) = \begin{cases} E(n_t, X_t) & \text{if } g(n_0, X_0, \ldots, n_{t-1}, X_{t-1}) = 1, \\ 0 & \text{if } g(n_0, X_0, \ldots, n_{t-1}, X_{t-1}) = 0, \end{cases} \tag{27}$$

where $E(\cdot, \cdot)$ is a fixed e-value and g is a decision rule determining whether the agent is permitted to conduct the next experiment. If $g(n_0, X_0, \ldots, n_{t-1}, X_{t-1}) = 0$ at any time $t-1$, the test process stops accumulating evidence and approval becomes impossible after time t.

Given a test process M, the utilities of both the agent and the principal can be defined analogously to Eq. 7, and a belief Markov decision process $\mathcal{M}^\varepsilon$ can be formulated as in Section 3, with the difference that the state space now includes any possible belief and any possible value of the test process, *i.e.*,[10]

$$\mathcal{S}^{\texttt{gen}} = \Delta(\Theta) \times \underbrace{\mathbb{R}_+}_{\text{total cost } C} \times \underbrace{\mathbb{R}_+}_{\text{test process } M} \cup S^{\texttt{out}},$$

[10] If the e-values depend on all past observations, as in Eq. 27, the state space may need to be augmented to explicitly track the history of experimental outcomes.

with a state S_t transitioning to S_{t+1} after taking action n_t and observing outcome X_t according to:

$$S_{t+1} = \begin{cases} (B_{t+1},\ C_t + c(n_t),\ M_{t+1}(n_t, X_t)) & \text{if } n_t \neq 0 \text{ and } 0 < f(S_t) < 1/\kappa, \\ S & \text{if } f(S_t) \geq 1/\kappa, \\ S^{\texttt{out}} & \text{if } n_t = 0 \text{ or } S_t = S^{\texttt{out}}, \end{cases} \tag{28}$$

where for $S_t = (B_t, C_t, M_t) \neq S^{\texttt{out}}$, $B_{t+1}(\theta) \propto B_t(\theta) \cdot P(X_t \mid n_t, \theta)$. In the above, f is the function that maps a state to the test process value, namely $f(S_t) = M_t$ for $S_t = (B_t, C_t, M_t) \neq S^{\texttt{out}}$ and $f(S^{\texttt{out}}) = 0$. While this formulation is fully general, it also significantly complicates the subsequent analysis of the optimal agent policy.

Alternative e-values and statistical power. For a given choice of test process M, it is important to note that the statistical power of the test—verifying whether $M_t \geq 1/\kappa$ at any time step—need not be 1. That is, with nonzero probability it may occur that, even for $\theta^* \geq \theta^{\texttt{b}}$, the approval process is unsuccessful.

To quantify the statistical power of sequential tests based on e-values (or supermartingales), there exists a canonical notion called *e-power*. For a given e-value E for the null H_0, its e-power against an alternative $L \in H_1$ is defined as $\mathbb{E}_L[\log E]$. Under this definition, likelihood ratios are the optimal e-values for simple nulls and alternatives, *i.e.*, when H_0 and H_1 each correspond to a single probability distribution. In contrast, when H_0 and H_1 are composite, as in the approval process of Section 4, the choice of e-value becomes more intricate (Waudby-Smith et al., 2025). For concreteness, in Section 4 we therefore adopt a general strategy that constructs e-values by exponentiating random variables (de la Peña, 1999; Waudby-Smith & Ramdas, 2022). The rationale is as follows: if we wish to test whether an arbitrary random variable X has mean $\mathbb{E}[X] < \theta^{\texttt{b}}$, we consider the quantity $X - \theta^{\texttt{b}}$, which we expect to be large when $\mathbb{E}[X] > \theta^{\texttt{b}}$ and small otherwise. To form an e-value, we consider the positive quantity $\exp(X - \theta^{\texttt{b}})$; since this can have expectation exceeding 1 even when $\mathbb{E}[X] < \theta^{\texttt{b}}$, we obtain a valid e-value satisfying Eq. 18 by shifting the argument by an appropriately chosen constant λ, *i.e.*, $\exp(X - \theta^{\texttt{b}} + \lambda)$, as given by Proposition 1. This e-value is particularly simple and admits a closed-form expression that simplifies the exposition in Section 3; however, in the context of the approval protocol in Section 2, other choices of e-values may achieve better statistical power for certain values of the efficacy parameter θ^*. Indeed, for E as defined in Proposition 1 and using the notation therein, the e-power is:

$$\mathbb{E}_{\theta^*}[\log E_t] = n_t \cdot \theta^* - n_t \cdot \log(1 + \theta^{\texttt{b}}(e-1)).$$

Thus, E_t has positive power $\mathbb{E}_{\theta^*}[\log E_t] \geq 0$ if and only if $\theta^* \geq \log(1 + \theta^{\texttt{b}}(e-1)) > \theta^{\texttt{b}}$. This means that if $\theta^{\texttt{b}} < \theta^* < \log(1 + \theta^{\texttt{b}}(e-1))$, it may be that $E_t \not\to \infty$ even as $n \to \infty$. In other words, the process M_t might never exceed the threshold $1/\kappa$ if θ^* is sufficiently close to—yet strictly above—$\theta^{\texttt{b}}$, even for RCTs of arbitrary size. The standard remedy (see Chapter 3 of (Ramdas & Wang, 2025)) is to form a *mixture* of e-values or supermartingales that generalizes Proposition 1. We now outline how our approval protocol can be extended to use such mixtures, and refer the reader to Appendix H.2.7 for additional experimental results.

Our starting point is the simple observation that the e-value $E(X_t, n_t)$ in Proposition 1 can be equivalently written as a likelihood ratio between a Binomial distribution evaluated at $\theta^{\texttt{b}}$ and one evaluated at a particular alternative:

$$E(X_t, n_t) = \left(\frac{\tilde{\theta}}{\theta^{\texttt{b}}}\right)^{X_t} \cdot \left(\frac{1-\tilde{\theta}}{1-\theta^{\texttt{b}}}\right)^{n_t - X_t},$$

where $\tilde{\theta} \geq \theta^{\texttt{b}}$ is the unique efficacy value satisfying:

$$e = \frac{\tilde{\theta} \cdot (1-\theta^{\texttt{b}})}{(1-\tilde{\theta}) \cdot \theta^{\texttt{b}}}.$$

To generalize this construction, one can mix over other values of the alternative parameter. This is achieved by choosing any (smooth) distribution $P^{\texttt{mix}}$ over $[\theta^{\texttt{b}}, 1]$ and defining the mixture test process as:[11]

$$M_{t+1}^{\texttt{mix}} = \int_{\theta^{\texttt{b}}}^{1} P^{\texttt{mix}}(\theta) \cdot \left(\frac{\theta}{\theta^{\texttt{b}}}\right)^{\sum_{i=0}^{t} X_i} \cdot \left(\frac{1-\theta}{1-\theta^{\texttt{b}}}\right)^{\sum_{i=0}^{t} n_i - \sum_{i=0}^{t} X_i} d\theta. \tag{29}$$

[11] An argument analogous to that of Proposition 1 shows that $M_t^{\texttt{mix}}$ is a supermartingale under H_0.

It can be shown that, if the true efficacy $\theta^* > \theta^\flat$ lies in the support of $P^{\texttt{mix}}$, then $M_{t+1}^{\texttt{mix}} \to \infty$ as $\sum_{i=0}^{t} n_i \to \infty$, *i.e.*, the mixture achieves asymptotic power 1 (Ramdas et al., 2023). In this context, Algorithm 3 can be used to compute the optimal agent policy for any subsidy with minimal modifications: it suffices to replace the test process function f with

$$f^{\texttt{mix}}(\alpha, \beta) = \int_{\theta^\flat}^{1} P^{\texttt{mix}}(\theta) \cdot \left(\frac{\theta}{\theta^\flat}\right)^{\alpha-\alpha_0} \cdot \left(\frac{1-\theta}{1-\theta^\flat}\right)^{\beta-\beta_0} d\theta,$$

and Algorithm 1 can then be used to compute the principal's optimal subsidy with no further modification.[12]

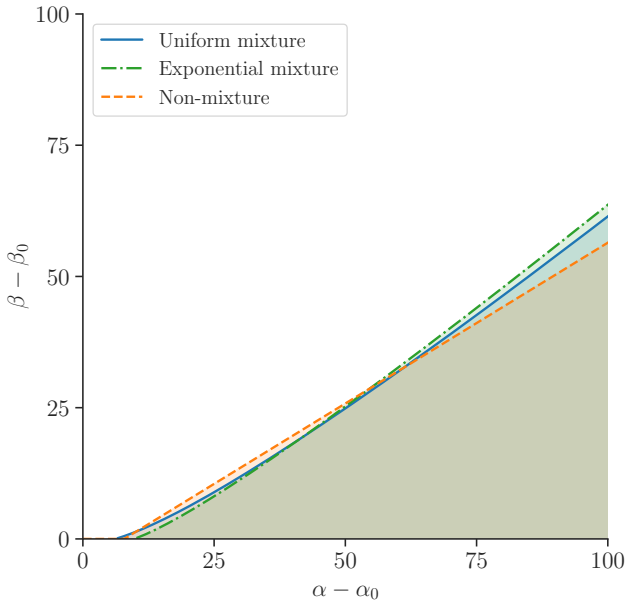


Figure 2. **Rejection regions in belief space.** The figure shows, for each agent belief with parameters (α, β), whether the condition $f(\alpha, \beta) \geq 1/\kappa$ is satisfied (*i.e.*, whether H_0 is rejected; shaded region), under different test processes: the test process defined using the non-mixed e-values in Proposition 1 (orange), a test process defined in Eq. 29 with a uniform mixture $P^{\texttt{mix}} = \mathrm{U}(\theta^\flat, 1)$ (blue), and a test process defined in Eq. 29 with an exponential mixture $P^{\texttt{mix}} = \mathrm{Exp}(10)$ restricted to $(\theta^\flat, 1)$ (green). Here, we set $\kappa = 0.05$, $\theta^\flat = 0.5$, and $\alpha_0 = \beta_0 = 1$.

We have presented the non-mixture process M defined in Section 2 in the main text for two reasons. First, it yields closed-form expressions which simplify the exposition and provide a more transparent (geometric) intuition for the belief Markov decision process (see Figure 3). Second, the multiplicative property in Eq. 3 is exploited extensively in the theoretical proofs in Appendix F, whereas the mixture process $M^{\texttt{mix}}$ does not satisfy this property in non-trivial cases where $P^{\texttt{mix}}$ is not a point mass. However, we conjecture that all our theoretical results in Section 3 and 4 carry over to the mixture setting—in particular, because the belief (α_t, β_t) remains a sufficient statistic for $M_t^{\texttt{mix}}$ via the monotone function $f^{\texttt{mix}}$, and the Bayesian update of the agent is unchanged—but adapting the proofs would require routing arguments through the Markov property of the belief MDP rather than through the algebraic identity $M_{t+1} = E_t \cdot M_t$. We leave a formal treatment of this extension for future work.

In Appendix H.2.7, we present experimental results using the above mixture test process with a uniform mixture $P^{\texttt{mix}} = \mathrm{U}(\theta^\flat, 1)$. Further, in Figure 2, we compare the regions of the belief space where $f(\alpha, \beta) \geq 1/\kappa$, *i.e.*, where H_0 is rejected, for different test processes. The test process without mixture (based on Proposition 1) yields a linear rejection region, while the mixture processes yield nonlinear regions whose boundary slope increases for larger values of (α, β).

[12]Note that the fact that Algorithm 1 returns an optimal subsidy only relies on the agent's optimal utility being a piecewise linear and convex function of the subsidy (Proposition 8), which holds independently of the specific form of f, as can be seen from its proof in Appendix F.10

D. Geometry of the Belief Markov Decision Process

In Figure 3, we illustrate the geometry of the belief MDP $\mathcal{M}^{\varepsilon}$ in the (α, β)-plane corresponding to all the possible beliefs for the agent. In blue, we represent the beliefs to which the agent may transition after selecting an action $n_t > 0$ (the figure shows $n_t = 3$ for concreteness). The exact next state depends on the realized outcome X_t of the experiment, as described in Eq. 8. Further, in light of Proposition 2, note that each pair (α, β) is associated with a value $f(\alpha, \beta)$ for the test process M, with pairs such that $f(\alpha, \beta) \geq 1/\kappa$ corresponding to a state where H_0 as been rejected by the principal. The condition

$$f(\alpha, \beta) \geq 1/\kappa \iff \alpha - \alpha_0 - (\alpha + \beta - \alpha_0 - \beta_0) \log(1 + \theta^{\flat}(e - 1)) \geq \log(1/\kappa) \tag{30}$$

corresponds to a linear region in the (α, β)-plane, represented in green in Figure 3 and labelled "Reject H_0". Note that the reward function defined in Eq. 9 only includes the positive term $\rho^{\mathbb{A}}$ for a transition that crosses the boundary of this region.

Similarly, the red region corresponds to beliefs for which—given a fixed total cost $C \geq 0$ and initial time step $l \in [T]$— it is optimal for the agent to opt out of the approval process by choosing $n = 0$. This region is bounded by the curve $\alpha \mapsto \tilde{\beta}(\alpha)$ in Proposition 6. Refer to Appendix H for concrete numerical solutions to the MDP $\mathcal{M}^{\varepsilon}$.

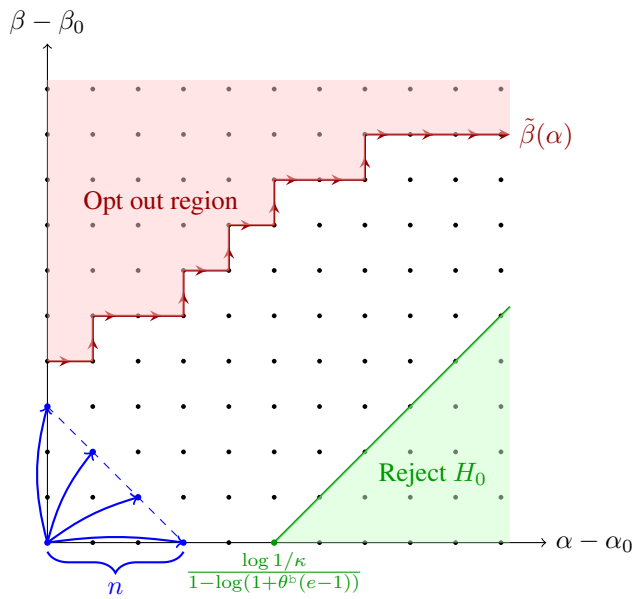


Figure 3. **Illustration of the geometry of the state space in the MDP $\mathcal{M}^{\varepsilon}$.**

E. Value Iteration in the Belief Markov Decision Process

In Algorithm 2, we present, for completeness, an adaptation of the value-iteration algorithm (Sutton & Barto, 2018) to exactly compute an optimal policy π^ε for the MDP $\mathcal{M}^\varepsilon$.[13]

In the proof of Proposition 4 in Appendix F.5, we showed that the set of accessible states in t time steps, which we denoted by $\mathcal{S}^{\mathrm{r}}(t)$, satisfies:

$$|\mathcal{S}^{\mathrm{r}}(t)| = \mathcal{O}\left((n^{\max})^2 \cdot t^2\right).$$

Building on this bound, it is straightforward to verify that the value-iteration method in Algorithm 2 has a time complexity $\mathcal{O}\left((n^{\max})^4 \cdot T^3\right)$. Indeed, for a given l, the set $|\mathcal{S}^{\mathrm{r}}(l) \setminus \{S^{\mathrm{out}}\}| = \mathcal{O}\left((n^{\max})^2 \cdot l^2\right)$, and for each such state, the inner loops (line 2 and 2) iterate over all actions n and sums over possible outcomes x. That is, for each l, the algorithm performs $\mathcal{O}\left((n^{\max})^4 \cdot l^2\right)$ iterations. Summing over l, we conclude that the total complexity is $\mathcal{O}\left((n^{\max})^4 \cdot T^3\right)$. We refer the reader to Appendix G for further details regarding the implementation of the algorithm.

Algorithm 2 Finds the Optimal Agent Policy π^ε for the belief MDP $\mathcal{M}^\varepsilon$

1: **Input** Subsidy ε, horizon T, max trials n^{max}, approval benefit ρ^A, cost function $c(n) = c_0 + c_1 n$, threshold κ, prior parameters (α_0, β_0), test process function f
2: **Initialize** $V^\varepsilon(S, T+1) \leftarrow 0$ **for all** states $S \in \mathcal{S}^{\mathrm{r}}$
3: **for** $l = T$ **down to** 0 **do**
4: **for** each $S = (\alpha, \beta, C) \in \mathcal{S}^{\mathrm{r}}(l) \setminus \{S^{\mathrm{out}}\}$ such that $f(\alpha, \beta) < 1/\kappa$ **do**
5: **for** each action $n \in \{1, \ldots, n^{max}\}$ **do**
6: $Q^\varepsilon(S, n, l) \leftarrow -(c_0 + c_1 n) + \sum_{x=0}^{n} \mathrm{BB}(n, \alpha, \beta)(x) \cdot \mathrm{NextValue}(x, n)$
7: **where** $\mathrm{NextValue}(x, n) =$
8: $\begin{cases} \rho^A + \varepsilon(C + c_0 + c_1 n) & \text{if } f(\alpha + x, \beta + n - x) \geq 1/\kappa \\ V^\varepsilon(\alpha + x, \beta + n - x, C + c_0 + c_1 n, l+1) & \text{otherwise} \end{cases}$
9: **end for**
10: $V^\varepsilon(S, l) \leftarrow \max\left\{0, \max_{n \in \{1, \ldots, n^{max}\}} Q^\varepsilon(S, n, l)\right\}$
11: **if** $V^\varepsilon(S, l) > 0$ **then**
12: $\pi^\varepsilon(S, l) \leftarrow \arg\max_{n \in \{1, \ldots, n^{max}\}} Q^\varepsilon(S, n, l)$
13: **else**
14: $\pi^\varepsilon(S, l) \leftarrow 0$
15: **end if**
16: **end for**
17: **end for**
18: **return** $V^\varepsilon, \pi^\varepsilon$

Note that the above algorithm can readily be modified to return the linear decomposition of the optimal value function at a subsidy ε in Proposition 7, namely $V^\varepsilon_{\pi^\varepsilon} = V^0_{\pi^\varepsilon} + \varepsilon \cdot A_{\pi^\varepsilon}$, as shown in Algorithm 3.

[13] $\mathrm{BB}(n, \alpha, \beta)$ denotes the Beta–Binomial distribution.

Algorithm 3 Finds the Optimal Agent Policy π^{ε} and Value Function Decomposition

Input Subsidy ε, horizon T, max trials n^{max}, approval benefit ρ^A, cost $c(n) = c_0 + c_1 n$, threshold κ, prior (α_0, β_0), test process function f
Initialize $V^{\varepsilon}(S, T+1) \leftarrow 0,\ V^0(S, T+1) \leftarrow 0,\ A(S, T+1) \leftarrow 0$ **for all** $S \in \mathcal{S}^{\mathrm{r}}$
for $l = T$ **down to** 0 **do**
  for each $S = (\alpha, \beta, C) \in \mathcal{S}^{\mathrm{r}}(l) \setminus \{S^{\mathtt{out}}\}$ s.t. $f(\alpha, \beta) < 1/\kappa$ **do**
    for each action $n \in \{1, \ldots, n^{max}\}$ **do**
      $c_n \leftarrow c_0 + c_1 n, \quad (\alpha'_x, \beta'_x) \leftarrow (\alpha{+}x,\ \beta{+}n{-}x)$
      $Q^{\varepsilon}(S, n, l) \leftarrow -c_n + \sum_{x=0}^{n} \mathrm{BB}(n, \alpha, \beta)(x) \cdot v^{\varepsilon}(x, n)$
      $Q^0(S, n, l) \leftarrow -c_n + \sum_{x=0}^{n} \mathrm{BB}(n, \alpha, \beta)(x) \cdot v^0(x, n)$
      $Q^A(S, n, l) \leftarrow \sum_{x=0}^{n} \mathrm{BB}(n, \alpha, \beta)(x) \cdot a(x, n)$
      where (writing **approved** for $f(\alpha'_x, \beta'_x) \geq 1/\kappa$):

$$v^{\varepsilon}(x, n) = \begin{cases} \rho^A + \varepsilon(C + c_n) & \text{if approved} \\ V^{\varepsilon}(\alpha'_x, \beta'_x, C{+}c_n, l{+}1) & \text{otherwise} \end{cases}$$

$$v^0(x, n) = \begin{cases} \rho^A & \text{if approved} \\ V^0(\alpha'_x, \beta'_x, C{+}c_n, l{+}1) & \text{otherwise} \end{cases}$$

$$a(x, n) = \begin{cases} C + c_n & \text{if approved} \\ A(\alpha'_x, \beta'_x, C{+}c_n, l{+}1) & \text{otherwise} \end{cases}$$

    end for
    $n^* \leftarrow \arg\max_{n \in \{1, \ldots, n^{max}\}} Q^{\varepsilon}(S, n, l)$
    if $Q^{\varepsilon}(S, n^*, l) > 0$ **then**
      $V^{\varepsilon}(S, l) \leftarrow Q^{\varepsilon}(S, n^*, l), \quad V^0(S, l) \leftarrow Q^0(S, n^*, l), \quad A(S, l) \leftarrow Q^A(S, n^*, l)$
      $\pi^{\varepsilon}(S, l) \leftarrow n^*$
    else
      $V^{\varepsilon}(S, l) \leftarrow 0, \quad V^0(S, l) \leftarrow 0, \quad A(S, l) \leftarrow 0$
      $\pi^{\varepsilon}(S, l) \leftarrow 0$
    end if
  end for
end for
return $V^{\varepsilon},\ V^0,\ A,\ \pi^{\varepsilon}$

F. Proofs

F.1. Proof of Proposition 1

Fix $\theta^* \in [0,1]$ and $n_t > 0$. Let $X_t \sim \mathrm{Bin}(n_t, \theta^*)$ and define

$$E(X_t, n_t) = \exp\left(X_t - n_t \cdot \log(1 + \theta^{\flat}(e-1))\right).$$

We explicitly show that the expectation of the above random variable is upper-bounded by 1 if $\theta^* \in H_0 = \left\{\theta^* : \theta^* < \theta^{\flat}\right\}$:

$$\begin{aligned}
\mathbb{E}_{X_t \sim \mathrm{Bin}(n_t, \theta^*)}[E(X_t, n_t)] &= \mathbb{E}_{X_t \sim \mathrm{Bin}(n_t, \theta^*)}[\exp\left(X_t - n_t \cdot \log(1 + \theta^{\flat}(e-1))\right)] \\
&= \frac{1}{(1 + \theta^{\flat}(e-1))^{n_t}} \cdot \mathbb{E}_{X_t \sim \mathrm{Bin}(n_t, \theta^*)}[\exp(X_t)] \\
&\overset{(*)}{=} \frac{1}{(1 + \theta^{\flat}(e-1))^{n_t}} \cdot (1 - \theta^* + \theta^* \cdot e)^{n_t} \\
&\overset{(**)}{\leq} \frac{1}{(1 + \theta^{\flat}(e-1))^{n_t}} \cdot (1 - \theta^{\flat} + \theta^{\flat} \cdot e)^{n_t} \\
&= 1
\end{aligned}$$

where in $(*)$ we have used the formula for the moment-generating function of the Binomial distribution, and in $(**)$ we have used that, by definition, if $\theta^* \in H_0$, then $\theta^* < \theta^{\flat}$.

F.2. Proof of Proposition 2

We begin by noting that the definition of the test process in Eq. 3 together with the e-value in Eq. 5 implies that, if the agent has continued the approval process up to time $t \leq T$ by selecting non-null $n_0, \ldots, n_t$, then:

$$M_{t+1} = \prod_{s=0}^{t} \exp\left(X_s - n_s \cdot \log(1 + \theta^{\mathrm{b}}(e-1))\right) \tag{31}$$

$$= \exp\left(\sum_{s=0}^{t} X_s - \sum_{s=0}^{t} n_s \cdot \log(1 + \theta^{\mathrm{b}}(e-1))\right) \tag{32}$$

Now, unfolding Eq. 1 for t time steps, we can write the parameters α_t and β_t that characterize the belief of the agent at time t as:

$$\begin{cases} \alpha_{t+1} = \alpha_0 + \sum_{s=0}^{t} X_s \\ \beta_{t+1} = \beta_0 + \sum_{s=0}^{t} n_s - \sum_{s=0}^{t} X_s, \end{cases} \tag{33}$$

$$\iff \begin{cases} \sum_{s=0}^{t} X_s = \alpha_{t+1} - \alpha_0 \\ \sum_{s=0}^{t} n_s = \beta_{t+1} - \beta_0 + \sum_{s=0}^{t} X_s = \alpha_{t+1} - \alpha_0 + \beta_{t+1} - \beta_0, \end{cases} \tag{34}$$

Thus, substituting the above in Eq. 31 we readily obtain:

$$\begin{aligned} M_{t+1} &= \exp\left(\sum_{s=0}^{t} X_s - \sum_{s=0}^{t} n_s \cdot \log(1 + \theta^{\mathrm{b}}(e-1))\right) \\ &= \exp\left(\alpha_{t+1} - \alpha_0 - (\alpha_{t+1} - \alpha_0 + \beta_{t+1} - \beta_0) \cdot \log(1 + \theta^{\mathrm{b}}(e-1))\right) \end{aligned}$$

F.3. Auxiliary Lemma to Proposition 3

Here, we prove a lemma that will be used later in the proof of Proposition 3.

Lemma 1. *Let $\pi \in \Pi$ be a policy, $S \in \mathcal{S}$ be a state such that $S \neq S^{out}$, denote $S = (\alpha, \beta, C)$ and fix an integrable function $G : \mathbb{R}^3 \to \mathbb{R}$. Further, denote by $\pi(S,t)(\bullet)$ the density over $\mathcal{A}$ given by π at time t, by $\mathrm{Beta}(\alpha,\beta)(\bullet)$ the density of the Beta distribution, by $\mathrm{Bin}(n,\theta)(\bullet)$ the density of a Binomial distribution with parameters n and θ, and by $\mathrm{BB}(n,\alpha,\beta)(\bullet)$ the density of a Beta-Binomial distribution. Consider any random variables θ^*, n_t and X_t with a joint density P such that n_t and θ^* are independent and:*[14]

$$\begin{cases} P(n_t) = \pi(S,t)(n_t) \\ P(\theta^*) = \mathrm{Beta}(\alpha,\beta)(\theta^*) \\ P(X_t \mid n_t, \theta^*) = \mathrm{Bin}(n_t, \theta^*)(X_t). \end{cases} \tag{35}$$

Then, it holds that:

$$\mathbb{E}_{\theta^* \sim \mathrm{Beta}(\alpha,\beta)} \mathbb{E}_{n_t \sim \pi(S,t)} \mathbb{E}_{X_t \sim \mathrm{Bin}(n_t,\theta^*)} [G(n_t, \theta^*, X_t)] = \mathbb{E}_{n_t \sim \pi(S,t)} \mathbb{E}_{X_t \sim \mathrm{BB}(n_t,\alpha,\beta)} \mathbb{E}_{\theta^* \sim \mathrm{Beta}(\alpha+X_t, \beta+n_t-X_t)} [G(n_t, \theta^*, X_t)]. \tag{36}$$

Proof. We first note that $X_t|n_t$ follows a Beta-Binomial distribution with parameters (n_t, α, β). Indeed, denoting by B the beta function and by Γ the gamma function:

$$\begin{aligned} P(X_t = k|n_t) &= \int_0^1 P(X_t = k|n_t, \theta^*) P(\theta^*)\, d\theta^* \\ &= \int_0^1 \left[\binom{n_t}{k} (\theta^*)^k (1-\theta^*)^{n_t-k} \right] \left[\frac{1}{B(\alpha,\beta)} (\theta^*)^{\alpha-1} (1-\theta^*)^{\beta-1} \right] d\theta^* \\ &= \frac{\binom{n_t}{k}}{B(\alpha,\beta)} \int_0^1 (\theta^*)^{k+\alpha-1} (1-\theta^*)^{n_t-k+\beta-1} d\theta^* \\ &= \binom{n_t}{k} \frac{B(k+\alpha, n_t-k+\beta)}{B(\alpha,\beta)} \\ &= \binom{n_t}{k} \frac{\Gamma(k+\alpha)\Gamma(n_t-k+\beta)}{\Gamma(n_t+\alpha+\beta)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}, \end{aligned}$$

which is precisely the density of a Beta-Binomial distribution with parameters (n_t, α, β). Similarly, $\theta^*|n_t, X_t$ follows a Beta distribution with parameters $(\alpha + X_t, \beta + n_t - X_t)$:

$$\begin{aligned} P(\theta^*|X_t, n_t) &= \frac{P(X_t|\theta^*, n_t) P(\theta^*)}{P(X_t|n_t)} \\ &= \frac{\left[\binom{n_t}{X_t} (\theta^*)^{X_t} (1-\theta^*)^{n_t-X_t} \right] \left[\frac{1}{B(\alpha,\beta)} (\theta^*)^{\alpha-1} (1-\theta^*)^{\beta-1} \right]}{\binom{n_t}{X_t} \frac{B(\alpha+X_t, \beta+n_t-Xt0)}{B(\alpha,\beta)}} \\ &= \frac{(\theta^*)^{X_t} (1-\theta^*)^{n_t-X_t} (\theta^*)^{\alpha-1} (1-\theta^*)^{\beta-1}}{B(\alpha+X_t, \beta+n_t-Xt0)} \\ &= \frac{(\theta^*)^{\alpha+X_t-1} (1-\theta^*)^{\beta+n_t-X_t-1}}{B(\alpha+X_t, \beta+n_t-X_t)}. \end{aligned}$$

Since n_t and θ^* are independent, the joint distribution can be written as:

$$P(\theta^*, n_t, X_t) = P(n_t) \cdot P(\theta^*) \cdot P(X_t|\theta^*, n_t) \tag{37}$$

Alternatively, we can factor the distribution P as:

$$P(\theta^*, n_t, X_t) = P(n_t) \cdot P(X_t|n_t) \cdot P(\theta^*|n_t, X_t). \tag{38}$$

[14] We adopt the notation θ^*, n_t, and X_t to align with the proof of Proposition 3, where this lemma is applied to the efficacy, sample size, and outcomes of the approval process.

Then, we can conclude using the law of total expectations for the expression $\mathbb{E}_{(n_t,\theta^*,X_t)\sim P}[G(n_t,\theta^*,X_t)]$ to obtain:

$$\mathbb{E}_{\theta^*\sim\mathrm{Beta}(\alpha,\beta)}\mathbb{E}_{n_t\sim\pi(S,t)}\mathbb{E}_{X_t\sim\mathrm{Bin}(n_t,\theta^*)}[G(n_t,\theta^*,X_t)] \\ = \mathbb{E}_{n_t\sim\pi(S,t)}\mathbb{E}_{X_t\sim\mathrm{BB}(n_t,\alpha,\beta)}\mathbb{E}_{\theta^*\sim\mathrm{Beta}(\alpha+X_t,\beta+n_t-X_t)}[G(n_t,\theta^*,X_t)] \quad (39)$$

□

F.4. Proof of Proposition 3

We fix a subsidy $\varepsilon \in [0, \varepsilon^{\texttt{max}}]$ and a policy $\pi \in \Pi$. Since ε is fixed, for clarity, we omit it from the notation for the remainder of the proof, and we will simply use the notation $r(\bullet)$ to denote the (agent's) reward in $\mathcal{M}^{\varepsilon}$. Similarly, V_π will denote the value function in $\mathcal{M}^{\varepsilon}$ for policy π. To show that Proposition 3 holds, we will proceed by induction over the time steps of the MDP $\mathcal{M}^{\varepsilon}$. We first introduce some additional notation.

Preliminaries. For any $t \leq T$ (representing a duration) and any $l \leq T - t$ (representing the initial time step), we define the following value function of π in the MDP $\mathcal{M}^{\varepsilon}$:

$$V_\pi(S, l, t) = \mathbb{E}_\pi \left[\sum_{k=l}^{l+t} r(S_k, n_k, S_{k+1}) \,\middle|\, S_l = S \right], \tag{40}$$

which is the total expected reward obtained by policy π starting from state S at time l and taking t steps.

Moreover, given a state $S = (\alpha, \beta, C) \in \mathcal{S} \setminus \{S^{\texttt{out}}\}$ of the MDP such that $f(S) < 1/\kappa$, $t \leq T$ and $l \leq T - t$, consider the approval process described in Section 2 when: (i) the initial time index is l, (ii) the agent has initial belief $B_0 = \mathrm{Beta}(\alpha, \beta)$, (iii) the test process M starts at $M_l = f(\alpha, \beta) < 1/\kappa$ and (iv) the subsidy includes an additional total cost C. Then, we define the agent's utility $\bar{U}^{\mathrm{A}}_{l,t}(\pi|S)$ (averaged over its initial belief) for the first t steps as:

$$\begin{aligned}
\bar{U}^{\mathrm{A}}_{l,t}(\pi|S) = \mathbb{E}_{\theta^* \sim \mathrm{Beta}(\alpha,\beta)} \Bigg[\mathbb{E}_{\pi_l} \Bigg[& \\
& \left(\rho^{\mathrm{A}} + \varepsilon \cdot \left(C + \sum_{j=0}^{\tau(S,l,t)} c(n_j) \right) \right) \cdot \mathbb{1}\{\exists j \in [\tau(S,l,t)] \colon f(\alpha,\beta) \cdot M_{j+1} \geq 1/\kappa\} \\
& - \left(\sum_{j=0}^{\tau(S,l,t)} c(n_j) \right) \Bigg| \theta^*, B_0 = \mathrm{Beta}(\alpha,\beta), C_0 = C \Bigg] \Bigg]
\end{aligned} \tag{41}$$

where $\tau(S, l, t) = t \wedge \min\{j \in \{0, \ldots, t\} \colon n_j = 0 \text{ or } f(\alpha, \beta) \cdot M_{j+1} \geq 1/\kappa\}$ is the last step of the approval process using the shifted policy π_l defined as $\pi_l(\bullet, j) = \pi(\bullet, j + l)$. Here, $\mathbb{E}_{\pi_l}[\bullet|\theta^*, B_0 = \mathrm{Beta}(\alpha, \beta), C_0 = C]$ indicates that: (i) all outcomes X_t are drawn with a fixed efficacy θ^*, and (ii) the agent has initial belief $\mathrm{Beta}(\alpha, \beta)$ and cumulated cost C. For instance, in the above expectation, the first action at index $j = 0$ taken by the agent is sampled from the distribution $\pi(\alpha, \beta, C, l)$. Alternatively, for a state S such that $S = S^{\texttt{out}}$ or $f(S) \geq 1/\kappa$, we simply define:

$$\bar{U}^{\mathrm{A}}_{l,t}(\pi|S) = 0. \tag{42}$$

Importantly, note that, by definition:

$$\begin{aligned}
& \bar{U}^{\mathrm{A}}_{0,T}(\pi \mid (\alpha_0, \beta_0, 0)) \\
=& \mathbb{E}_{\theta^* \sim \mathrm{Beta}(\alpha_0,\beta_0)} \Bigg[\mathbb{E}_{\pi_0} \Bigg[\left(\rho^{\mathrm{A}} + \varepsilon \sum_{j=0}^{\tau(\alpha_0,\beta_0,0,0,0)} c(n_j) \right) \cdot \mathbb{1}\{\exists j \in [\tau(\alpha_0,\beta_0,0,0,0)] : M_{j+1} \geq 1/\kappa\} \\
& - \sum_{j=0}^{\tau(\alpha_0,\beta_0,0,0,0)} c(n_j) \Bigg| \theta^*, B_0 = \mathrm{Beta}(\alpha_0,\beta_0), C_0 = 0 \Bigg] \Bigg], \\
=& \mathbb{E}_{\theta^* \sim B_0} \big[U^{\mathrm{A}}(\pi, \varepsilon) \big].
\end{aligned}$$

In light of the above, our goal will be to show by induction over t that:

$$\boxed{\bar{U}^{\mathrm{A}}_{l,t}(\pi \mid S) = V_\pi(S, l, t) \text{ for any } S \in \mathcal{S},\ t \leq T,\ l \leq T - t.}$$

Then, by particularizing to $t = T$, $l = 0$ and $S = (\alpha_0, \beta_0, 0)$ and using Eq. 10, we will obtain that

$$\bar{U}^{\mathrm{A}}(\pi; \varepsilon) = V_\pi(\alpha_0, \beta_0, 0, 0, T) = \bar{U}^{\mathrm{A}}_{0,T}(\pi|\, \alpha_0, \beta_0, 0\,) = \mathbb{E}_{\theta^* \sim B_0}[U^{\mathrm{A}}(\pi, \varepsilon)],$$

which is the statement in Proposition 3.

Base case $t = 0$**.** We will show that $\bar{U}^{\mathrm{A}}_{l,0}(\pi|S) = V_\pi(S, l, 0)$ holds for any $S \in \mathcal{S}$ and $l \leq T$. We first consider the case where $S = S^{\mathtt{out}}$ or $f(S) \geq 1/\kappa$. In this case, note that the transition dynamics of the MDP (Eq. 8) imply that at any time step $k \geq l$, $S_k = S^{\mathtt{out}}$ or $S_k = S$. Thus, the expression in Eq. 40 only contains rewards that are null (see Eq. 9), and hence:

$$V_\pi(S, l, 0) = 0.$$

Similarly, by definition (Eq. 42), $\bar{U}^{\mathrm{A}}_{l,0}(\pi|S) = 0$. Thus, $\bar{U}^{\mathrm{A}}_{l,0}(\pi|S) = V_\pi(S, l, 0)$.

Consider now the non-trivial case where $S \neq S^{\mathtt{out}}$ and $f(S) < 1/\kappa$, and write $S = (\alpha, \beta, C)$. In this case, the value function starting at time l for $t = 0$ can be expanded as (note that $t = 0$ corresponds to a single step in the MDP):

$$\begin{aligned}
V_\pi(S, l, 0) &= \mathbb{E}_\pi \left[r(S_l, n_l, S_{l+1}) | S_l = S \right] \\
&= \mathbb{E}_\pi \left[-c(n_l) + (\rho^A + \varepsilon \cdot (C + c(n_l))) \cdot \mathbb{1}\{f(S_{l+1}) \geq 1/\kappa)\} \middle| S_l = S \right] \\
&= \mathbb{E}_{n_l \sim \pi(S,l)} \mathbb{E}_{X_l \sim \mathrm{BB}(n_l, \alpha, \beta)} \big[-c(n_l) \\
&\quad + (\rho^A + \varepsilon \cdot (C + c(n_l))) \cdot \mathbb{1}\{f(S) \cdot E(X_l, n_l) \geq 1/\kappa)\} \big]
\end{aligned}$$

In the above, we denote by $X_l \sim \mathrm{BB}(n_l, \alpha, \beta)$ a sample from the Beta-Binomial distribution, corresponding to sampling $\theta_l \sim \mathrm{Beta}(\alpha, \beta)$, and then $X_l \sim \mathrm{Bin}(n_l, \theta_l)$, and we have used the definition of the test process in Eq. 3. On the other hand, let $t = 0$ in Eq. 41. Then, $\tau(S, l, 0) = 0$, and we obtain:[15]

$$\begin{aligned}
&\bar{U}^{\mathrm{A}}_{l,0}(\pi|S) \\
&= \mathbb{E}_{\theta^* \sim \mathrm{Beta}(\alpha,\beta)} \Bigg[\mathbb{E}_{\pi_l} \Bigg[-c(n_0) + (\rho^{\mathrm{A}} + \varepsilon \cdot (C + c(n_0))) \cdot \mathbb{1}\{f(\alpha, \beta) \cdot M_1 \geq 1/\kappa\} \\
&\quad \Bigg| \theta^*, B_0 = \mathrm{Beta}(\alpha, \beta), C_0 = C \Bigg] \Bigg] \\
&= \mathbb{E}_{\theta^* \sim \mathrm{Beta}(\alpha,\beta)} \Bigg[\mathbb{E}_{\pi_l} \Bigg[-c(n_0) + (\rho^{\mathrm{A}} + \varepsilon \cdot (C + c(n_0))) \cdot \mathbb{1}\{f(S) \cdot M_1 \geq 1/\kappa\} \\
&\quad \Bigg| \theta^*, B_0 = \mathrm{Beta}(\alpha, \beta), C_0 = C \Bigg] \Bigg] \\
&= \mathbb{E}_{\theta^* \sim \mathrm{Beta}(\alpha,\beta)} \mathbb{E}_{n_0 \sim \pi(S,l)} \mathbb{E}_{X_0 \sim \mathrm{Bin}(n_0, \theta^*)} \Bigg[-c(n_0) \\
&\quad + (\rho^{\mathrm{A}} + \varepsilon \cdot (C + c(n_0))) \cdot \mathbb{1}\{f(S) \cdot M_1 \geq 1/\kappa\} \Bigg] \\
&\overset{(*)}{=} \mathbb{E}_{n_0 \sim \pi(S,l)} \mathbb{E}_{\theta^* \sim \mathrm{Beta}(\alpha,\beta)} \mathbb{E}_{X_0 \sim \mathrm{Bin}(n_0, \theta^*)} \Bigg[-c(n_0) \\
&\quad + (\rho^{\mathrm{A}} + \varepsilon \cdot (C + c(n_0))) \cdot \mathbb{1}\{f(S) \cdot E(n_0, X_0) \geq 1/\kappa\} \Bigg] \\
&= \mathbb{E}_{n_l \sim \pi(S,l)} \mathbb{E}_{X_l \sim \mathrm{BB}(n_l, \alpha, \beta)} \Bigg[-c(n_l) + (\rho^{\mathrm{A}} + \varepsilon \cdot (C + c(n_l))) \cdot \mathbb{1}\{f(S) \cdot E(n_l, X_l) \geq 1/\kappa\} \Bigg] \\
&= V_\pi(S, l, 0),
\end{aligned}$$

[15] Whenever we write nested expectations, such as $\mathbb{E}_Y \mathbb{E}_Z[\bullet]$ for arbitrary random variables Y and Z, the inner expectation is understood to be conditional on the outer variable; that is, $\mathbb{E}_Y \left[\mathbb{E}_Z[\bullet \mid Y] \right]$. To simplify notation, we may omit the explicit conditioning when no confusion is likely to arise.

where in $(*)$ we have used that θ^* and n_0 are independent (since n_0 is sampled independently from $\pi(S, l)$). This concludes the base case.

Inductive step $t \to t+1$**.** For the inductive step, assume that given a $t < T$, for any $l \leq T - t$ and for any state $S \in \mathcal{S}$ it holds that:

$$\bar{U}^{\mathbb{A}}_{l,t}(\pi \mid S) = V_\pi(S, l, t).$$

We will show that $\bar{U}^{\mathbb{A}}_{l,t+1}(\pi|S) = V_\pi(S, l, t+1)$ for any $l \leq T - (t+1)$ and $S \in \mathcal{S}$. To this end, fix any such S and l.

Firstly, we consider the case where $S = S^{\text{out}}$ or $f(S) \geq 1/\kappa$, where the transition dynamics of the MDP (Eq. 8) imply that at any time step $k \geq l$, $S_k = S^{\text{out}}$ or $S_k = S$. Then, similarly to the base case, $V_\pi(S, l, t) = 0$ and $\bar{U}^{\mathbb{A}}_{l,t}(\pi|S) = 0$ for any $t \geq 0$, so the equality holds.

We focus now on the non-trivial case where $S \neq S^{\text{out}}$ and $f(S) < 1/\kappa$, and write $S = (\alpha, \beta, C)$. We begin by expanding the value for the first $t+1$ time steps in Eq. 40:

$$\begin{aligned}
V_\pi(S, l, t+1) &= \mathbb{E}_\pi \left[\sum_{k=l}^{l+t+1} r(S_k, n_k, S_{k+1}) \,\middle|\, S_l = S \right] \\
&= \mathbb{E}_\pi \left[r(S_l, n_l, S_{l+1}) + \sum_{k=l+1}^{l+t+1} r(S_k, n_k, S_{k+1}) \,\middle|\, S_l = S \right] \\
&= \mathbb{E}_\pi \left[r(S_l, n_l, S_{l+1}) + V_\pi(S_{l+1}, l+1, t) \,|\, S_l = S \right] \\
&\overset{(*)}{=} \mathbb{E}_\pi \left[r(S_l, n_l, S_{l+1}) + \bar{U}^{\mathbb{A}}_{l+1,t}(\pi|S_{l+1}) \,\middle|\, S_l = S \right] \\
&\overset{(**)}{=} \underbrace{\mathbb{E}_\pi \left[r(S_l, n_l, S_{l+1}) \,|\, S_l = S \right]}_{\dagger} + \underbrace{\mathbb{E}_\pi \left[\bar{U}^{\mathbb{A}}_{l+1,t}(\pi|S_{l+1}) \cdot \mathbb{1}\{f(S_{l+1}) < 1/\kappa\} \,\middle|\, S_l = S \right]}_{\ddagger}
\end{aligned} \tag{43}$$

where $(*)$ follows from the induction hypothesis and $(**)$ because $\bar{U}^{\mathbb{A}}_t(\pi|S_{l+1}) = 0$ if $f(S_{l+1}) \geq 1/\kappa$. Analogously to the base case, the term $\dagger$ above can be written as:

$$\begin{aligned}
\dagger &= \mathbb{E}_\pi \left[r(S_l, n_l, S_{l+1}) | S_l = S \right] \\
&= \mathbb{E}_\pi \left[-c(n_l) + (\rho^{\mathbb{A}} + \varepsilon \cdot (C + c(n_l))) \cdot \mathbb{1}\{f(S_{l+1}) \geq 1/\kappa\} | S_l = S \right] \\
&= \mathbb{E}_{n_l \sim \pi(S,l)} \mathbb{E}_{\theta^* \sim \text{Beta}(\alpha,\beta)} \mathbb{E}_{X_l \sim \text{Bin}(n_l, \theta^*)} \Big[-c(n_l) \\
&\quad + (\rho^{\mathbb{A}} + \varepsilon \cdot (C + c(n_l))) \cdot \mathbb{1}\{f(S_{l+1}) \geq 1/\kappa\} \Big] \\
&= \mathbb{E}_{\theta^* \sim \text{Beta}(\alpha,\beta)} \mathbb{E}_{n_l \sim \pi(S,l)} \mathbb{E}_{X_l \sim \text{Bin}(n_l, \theta^*)} \Big[-c(n_l) \\
&\quad + (\rho^{\mathbb{A}} + \varepsilon \cdot (C + c(n_l))) \cdot \mathbb{1}\{f(\alpha + X_l, \beta + n_l - X_l) \geq 1/\kappa\} \Big] \\
&= \mathbb{E}_{\theta^* \sim \text{Beta}(\alpha,\beta)} \mathbb{E}_{n_0 \sim \pi_l(S,0)} \mathbb{E}_{X_0 \sim \text{Bin}(n_0, \theta^*)} \Big[-c(n_0) \\
&\quad + (\rho^{\mathbb{A}} + \varepsilon \cdot (C + c(n_0))) \cdot \mathbb{1}\{f(\alpha + X_0, \beta + n_0 - X_0) \geq 1/\kappa\} \Big] \\
&= \mathbb{E}_{\theta^* \sim \text{Beta}(\alpha,\beta)} \mathbb{E}_{n_0 \sim \pi_l(S,0)} \mathbb{E}_{X_0 \sim \text{Bin}(n_0, \theta^*)} \Big[-c(n_0) \\
&\quad + (\rho^{\mathbb{A}} + \varepsilon \cdot (C + c(n_0))) \cdot \mathbb{1}\{f(\alpha, \beta) \cdot E(n_0, X_0) \geq 1/\kappa\} \Big] \\
&= \mathbb{E}_{\theta^* \sim \text{Beta}(\alpha,\beta)} \Big[\mathbb{E}_{\pi_l} \Big[-c(n_0) + (\rho^{\mathbb{A}} + \varepsilon \cdot (C + c(n_0))) \cdot \mathbb{1}\{f(\alpha, \beta) \cdot M_1 \geq 1/\kappa\} \\
&\quad \Big| \theta^*, B_0 = \text{Beta}(\alpha, \beta), C_0 = C \Big] \Big]
\end{aligned} \tag{44}$$

We now focus on the term $\ddagger$ in Eq. 43. Given $S_l = S$, the state S_{l+1} is fully determined by the action n_l and the value X_l,

namely (see Eq. 8),

$$S_{l+1} = \begin{cases} (\alpha + X_l, \beta + n_l - X_l, C + c(n_l)) & \text{if} \quad n_l > 0 \\ S^{\texttt{out}} & \text{if} \quad n_l = 0. \end{cases}$$

Thus, conditional on the policy $\pi(S, l)$ selecting action $n_l = 0$, we have $S_{l+1} = S^{\texttt{out}}$ and $\bar{U}^{\mathbb{A}}_{l+1,t}(\pi|S_{l+1}) = 0$. That is, conditioning on the the event $\{n_l = 0\}$ inside the expectations † and ‡ results in:

$$\begin{cases} \mathbb{E}_\pi \left[r(S_l, n_l, S_{l+1}) \,|\, S_l = S, n_l = 0 \right] = 0 \\ \mathbb{E}_\pi \left[\bar{U}^{\mathbb{A}}_{l+1,t}(\pi|S_{l+1}) \cdot \mathbb{1}\{f(S_{l+1}) < 1/\kappa\} \,\middle|\, S_l = S, n_l = 0 \right] = 0. \end{cases}$$

Similarly, note that conditional on the policy $\pi(\alpha, \beta, C, l)$ selecting action $n_l = 0$, we also have:

$$\mathbb{E}_{\theta^* \sim \text{Beta}(\alpha,\beta)} \Bigg[\mathbb{E}_{\pi_l} \Bigg[\left(\rho^{\mathbb{A}} + \varepsilon \cdot \left(C + \sum_{j=0}^{\tau(S,l,t+1))} c(n_j) \right) \right) \cdot \mathbb{1}\{\exists j \in [\tau(S, l, t+1)] \colon f(\alpha, \beta) \cdot M_{j+1} \geq 1/\kappa\} - \left(\sum_{j=0}^{\tau(S,l,t+1)} c(n_j) \right) \Bigg| \theta^*, B_0 = \text{Beta}(\alpha, \beta), C_0 = C, n_0 = 0 \Bigg] \Bigg] = 0$$

As a consequence, in the following, we assume without loss of generality that the action 0 is not in the support of the distribution $\pi(S, l)$, and therefore $S_{l+1} \neq S^{\texttt{out}}$.

Then, under the above simplification, we expand the expectation in the term ‡ to average over the possible values of n_l and X_l, and then substitute the expression for the utility $\bar{U}^{\mathbb{A}}_{l+1,t}(\pi|S_{l+1})$ defined in Eq. 41 (emphasized below in blue for clarity):[16]

$$\begin{aligned} \ddagger &= \mathbb{E}_\pi \left[\bar{U}^{\mathbb{A}}_{l+1,t}(\pi|S_{l+1}) \cdot \mathbb{1}\{f(S_{l+1}) < 1/\kappa\} \,\middle|\, S_l = S \right] \\ &= \mathbb{E}_{n'_l \sim \pi(S,l)} \mathbb{E}_{X'_l \sim \text{BB}(n'_l, \alpha, \beta)} \mathbb{E}_{\theta^* \sim \text{Beta}(\alpha + X'_l, \beta + n'_l - X'_l)} \Bigg[\mathbb{E}_{\pi_{l+1}} \Bigg[\mathbb{1}\{f(\alpha + X'_l, \beta + n'_l - X'_l) < 1/\kappa\} \cdot \\ &\left(\rho^{\mathbb{A}} + \varepsilon \cdot \left(C + c(n'_l) + \sum_{j=0}^{\tilde{\tau}} c(n_j) \right) \right) \cdot \mathbb{1}\{\exists j \in [\tilde{\tau}] \colon f(\alpha + X'_l, \beta + n'_l - X'_l) \cdot M_{j+1} \geq 1/\kappa\} \\ &- \mathbb{1}\{f(\alpha + X'_l, \beta + n'_l - X'_l) < 1/\kappa\} \cdot \left(\sum_{j=0}^{\tilde{\tau}} c(n_j) \right) \\ &\Bigg| \theta^*, B_0 = \text{Beta}(\alpha + X'_l, \beta + n'_l - X'_l), C_0 = C + c(n'_l) \Bigg] \Bigg], \end{aligned}$$

where we have defined $\tilde{\tau} = \tau(S(n'_l, X'_l), l + 1, t)$, and we use $S(n'_l, X'_l)$ to denote the state to which S transitions after selecting action n'_l and observing the outcome X'_l, as given by the transition dynamics (Eq. 8).

[16] We also change notation $n_l \to n'_l$ and $X_l \to X'_l$ to avoid confusion with the actions and outcomes that appear when expanding $\bar{U}^{\mathbb{A}}_{l+1,t}(\pi|S_{l+1})$.

We now leverage Lemma 1 to reorder the first three expectations as:

$$
\begin{aligned}
\ddagger &= \mathbb{E}_{\theta^* \sim \mathrm{Beta}(\alpha,\beta)} \mathbb{E}_{n_l' \sim \pi(S,l)} \mathbb{E}_{X_l' \sim \mathrm{Bin}(n_l', \theta^*)} \Bigg[\mathbb{E}_{\pi_{l+1}} \Bigg[\mathbb{1}\{f(\alpha + X_l', \beta + n_l' - X_l') < 1/\kappa\} \cdot \\
&\left(\rho^{\mathrm{A}} + \varepsilon \cdot \left(C + c(n_l') + \sum_{j=0}^{\tilde{\tau}} c(n_j) \right) \right) \cdot \mathbb{1}\{\exists j \in [\tilde{\tau}] \colon f(\alpha + X_l', \beta + n_l' - X_l') \cdot M_{j+1} \geq 1/\kappa\} \\
&- \mathbb{1}\{f(\alpha + X_l', \beta + n_l' - X_l') < 1/\kappa\} \cdot \left(\sum_{j=0}^{\tilde{\tau}} c(n_j) \right) \\
&\Bigg| \theta^*, B_0 = \mathrm{Beta}(\alpha + X_l', \beta + n_l' - X_l'), C_0 = C + c(n_l') \Bigg] \Bigg] \\
&= \mathbb{E}_{\theta^* \sim \mathrm{Beta}(\alpha,\beta)} \mathbb{E}_{n_0 \sim \pi(S,l)} \mathbb{E}_{X_0 \sim \mathrm{Bin}(n_0, \theta^*)} \Bigg[\mathbb{1}\{f(\alpha + X_0, \beta + n_0 - X_0) < 1/\kappa\} \cdot \mathbb{E}_{\pi_{l+1}} \Bigg[\\
&\left(\rho^{\mathrm{A}} + \varepsilon \cdot \left(C + c(n_0) + \sum_{j=0}^{\tilde{\tau}} c(\tilde{n}_j) \right) \right) \cdot \mathbb{1}\{\exists j \in [\tilde{\tau}] \colon f(\alpha + X_0, \beta + n_0 - X_0) \cdot \tilde{M}_{j+1} \geq 1/\kappa\} \\
&- \left(\sum_{j=0}^{\tilde{\tau}} c(\tilde{n}_j) \right) \Bigg] \Bigg| \theta^*, B_0 = \mathrm{Beta}(\alpha + X_0, \beta + n_0 - X_0), C_0 = C + c(n_0) \Bigg],
\end{aligned}
\tag{45}
$$

where, in the last equality, we have renamed the dummy variables appearing in the inner expectation using tildes, *i.e.*, $\tilde{n}_j$, $\tilde{X}_j$, and $\tilde{M}_j$.[17] This is purely notational at this stage, but will be useful later in the proof when it becomes important to distinguish between different sets of variables.

To keep our objective in view, recall that the goal of the inductive step is to show that

$$\bar{U}^{\mathrm{A}}_{l,t+1}(\pi \mid S) = \dagger + \ddagger.$$

Our goal will now be to expand the term $\bar{U}^{\mathrm{A}}_{l,t+1}(\pi \mid S)$ and verify that this equality indeed holds. To this end, we will use the identity

$$1 = \mathbb{1}\{f(\alpha, \beta) \cdot M_1 \geq 1/\kappa\} + \mathbb{1}\{f(\alpha, \beta) \cdot M_1 < 1/\kappa\},$$

which holds for any α, β and M_1 since the two events in the indicator functions are complementary, and substitute it in the definition of $\bar{U}^{\mathrm{A}}_{l,t+1}(\pi \mid S)$ in Eq. 41:

[17] Here, note that $\tilde{\tau}(S(n_0, X_0), l+1, t) \stackrel{d}{=} \tilde{\tau}(S(n_l', X_l'), l+1, t)$, because $n_0 \stackrel{d}{=} n_l'$ and $X_0 \stackrel{d}{=} X_l'$ by definition, and so we kept the notation $\tilde{\tau}$ to denote $\tau(S(n_0, X_0), l+1, t)$

$$
\begin{aligned}
&\bar{U}^{\mathbb{A}}_{l,t+1}(\pi|S) \\
&= \mathbb{E}_{\theta^* \sim \mathrm{Beta}(\alpha,\beta)}\Bigg[\mathbb{E}_{\pi_l}\Bigg[\left(\rho^{\mathbb{A}} + \varepsilon \cdot \left(C + \sum_{j=0}^{\tau(S,l,t+1)} c(n_j)\right)\right) \\
&\quad \cdot \mathbb{1}\{\exists j \in [\tau(S,l,t+1)] \colon f(\alpha,\beta) \cdot M_{j+1} \geq 1/\kappa\} \\
&\quad - \left(\sum_{j=0}^{\tau(S,l,t+1)} c(n_j)\right) \Bigg| \theta^*, B_0 = \mathrm{Beta}(\alpha,\beta), C_0 = C\Bigg]\Bigg] \\
&= \mathbb{E}_{\theta^* \sim \mathrm{Beta}(\alpha,\beta)}\Bigg[\mathbb{E}_{\pi_l}\Bigg[\left(\rho^{\mathbb{A}} + \varepsilon \cdot \left(C + \sum_{j=0}^{\tau(S,l,t+1)} c(n_j)\right)\right) \\
&\quad \cdot \mathbb{1}\{f(\alpha,\beta) \cdot M_1 \geq 1/\kappa\} \cdot \mathbb{1}\{\exists j \in [\tau(S,l,t+1)] \colon f(\alpha,\beta) \cdot M_{j+1} \geq 1/\kappa\} \\
&\quad - \mathbb{1}\{f(\alpha,\beta) \cdot M_1 \geq 1/\kappa\} \cdot \left(\sum_{j=0}^{\tau(S,l,t+1)} c(n_j)\right) \Bigg| \theta^*, B_0 = \mathrm{Beta}(\alpha,\beta), C_0 = C\Bigg]\Bigg] \\
&\quad + \mathbb{E}_{\theta^* \sim \mathrm{Beta}(\alpha,\beta)}\Bigg[\mathbb{E}_{\pi_l}\Bigg[\left(\rho^{\mathbb{A}} + \varepsilon \cdot \left(C + \sum_{j=0}^{\tau(S,l,t+1)} c(n_j)\right)\right) \\
&\quad \cdot \mathbb{1}\{f(\alpha,\beta) \cdot M_1 < 1/\kappa\} \cdot \mathbb{1}\{\exists j \in [\tau(S,l,t+1)] \colon f(\alpha,\beta) \cdot M_{j+1} \geq 1/\kappa\} \\
&\quad - \mathbb{1}\{f(\alpha,\beta) \cdot M_1 < 1/\kappa\} \cdot \left(\sum_{j=0}^{\tau(S,l,t+1)} c(n_j)\right) \Bigg| \theta^*, B_0 = \mathrm{Beta}(\alpha,\beta), C_0 = C\Bigg]\Bigg]
\end{aligned}
$$

We now observe that if $f(\alpha,\beta) \cdot M_1 \geq 1/\kappa$, then $\tau(S,l,t+1) = 0$, that is, the approval process stops after the first step, which simplifies the first summand and yields:

$$
\begin{aligned}
&\bar{U}^{\mathbb{A}}_{l,t+1}(\pi|S) \\
&= \mathbb{E}_{\theta^* \sim \mathrm{Beta}(\alpha,\beta)}\Bigg[\mathbb{E}_{\pi_l}\Bigg[-c(n_0) + (\rho^{\mathbb{A}} + \varepsilon \cdot (C + c(n_0))) \cdot \mathbb{1}\{f(\alpha,\beta) \cdot M_1 \geq 1/\kappa\} \\
&\qquad\qquad \Bigg| \theta^*, B_0 = \mathrm{Beta}(\alpha,\beta), C_0 = C\Bigg]\Bigg] \\
&\quad + \mathbb{E}_{\theta^* \sim \mathrm{Beta}(\alpha,\beta)}\Bigg[\mathbb{E}_{\pi_l}\Bigg[\left(\rho^{\mathbb{A}} + \varepsilon \cdot \left(C + \sum_{j=0}^{\tau(S,l,t+1)} c(n_j)\right)\right) \\
&\quad \cdot \mathbb{1}\{f(\alpha,\beta) \cdot M_1 < 1/\kappa\} \cdot \mathbb{1}\{\exists j \in [\tau(S,l,t+1)] \colon f(\alpha,\beta) \cdot M_{j+1} \geq 1/\kappa\} \\
&\quad - \mathbb{1}\{f(\alpha,\beta) \cdot M_1 < 1/\kappa\} \cdot \left(\sum_{j=1}^{\tau(S,l,t+1)} c(n_j)\right) \Bigg| \theta^*, B_0 = \mathrm{Beta}(\alpha,\beta), C_0 = C\Bigg]\Bigg]
\end{aligned}
$$

Further, we can identify that the first summand in the above expression corresponds to the form of the term † derived in Eq. 44, and replacing it, we obtain:

$$
\begin{aligned}
&\bar{U}^{\mathbb{A}}_{l,t+1}(\pi|S) \\
&= \dagger \\
&\quad + \mathbb{E}_{\theta^* \sim \mathrm{Beta}(\alpha,\beta)}\Bigg[\mathbb{E}_{\pi_l}\Bigg[\left(\rho^{\mathbb{A}} + \varepsilon \cdot \left(C + \sum_{j=0}^{\tau(S,l,t+1)} c(n_j)\right)\right) \\
&\quad \cdot \mathbb{1}\{f(\alpha,\beta)\cdot M_1 < 1/\kappa\} \cdot \mathbb{1}\{\exists j \in [\tau(S,l,t+1)] \colon f(\alpha,\beta)\cdot M_{j+1} \geq 1/\kappa\} \\
&\quad - \mathbb{1}\{f(\alpha,\beta)\cdot M_1 < 1/\kappa\} \cdot \left(\sum_{j=1}^{\tau(S,l,t+1)} c(n_j)\right) \Bigg| \theta^*, B_0 = \mathrm{Beta}(\alpha,\beta), C_0 = C\Bigg]\Bigg]
\end{aligned}
$$

We now focus on the orange term in the expression above, which we denote by $\square$. Our goal will be to show that $\square = \ddagger$. We start by factoring out the term $\mathbb{1}\{f(\alpha,\beta)\cdot M_1 < 1/\kappa\}$, and using that by definition $M_1 = E(n_0, M_0)$, we obtain:

$$
\begin{aligned}
\square &= \mathbb{E}_{\theta^* \sim \mathrm{Beta}(\alpha,\beta)}\Bigg[\mathbb{E}_{\pi_l}\Bigg[\mathbb{1}\{f(\alpha,\beta)\cdot E(n_0, X_0) < 1/\kappa\} \cdot \Bigg(\\
&\quad \left(\rho^{\mathbb{A}} + \varepsilon \cdot \left(C + c(n_0) + \sum_{j=1}^{\tau(S,l,t+1)} c(n_j)\right)\right) \mathbb{1}\{\exists j \in [\tau(S,l,t+1)] \colon f(\alpha,\beta)\cdot M_{j+1} \geq 1/\kappa\} \\
&\quad - \sum_{j=1}^{\tau(S,l,t+1)} c(n_j)\Bigg) \Bigg| \theta^*, B_0 = \mathrm{Beta}(\alpha,\beta), C_0 = C\Bigg]\Bigg]
\end{aligned}
$$

In the above expectation $\mathbb{E}_{\pi_l}[\bullet|\theta^*, B_0 = \mathrm{Beta}(\alpha,\beta), C_0 = C]$, we can use the law of iterated expectations by conditioning on the first samples n_0 and X_0. In particular, note that $n_0 \sim \pi(\alpha,\beta,C,l)$ and that $X_0 \sim \mathrm{Bin}(n_0,\theta^*)$, and thus:

$$
\begin{aligned}
\square &= \mathbb{E}_{\theta^* \sim \mathrm{Beta}(\alpha,\beta)}\mathbb{E}_{n_0 \sim \pi(S,l)}\mathbb{E}_{X_0 \sim \mathrm{Bin}(n_0,\theta^*)}\Bigg[\mathbb{E}_{\pi_l}\Bigg[\mathbb{1}\{f(\alpha,\beta)\cdot E(n_0, X_0) < 1/\kappa\} \cdot \Bigg(\\
&\quad \left(\rho^{\mathbb{A}} + \varepsilon \cdot \left(C + c(n_0) + \sum_{j=1}^{\tau(S,l,t+1)} c(n_j)\right)\right) \mathbb{1}\{\exists j \in [\tau(S,l,t+1)] \colon f(\alpha,\beta)\cdot M_{j+1} \geq 1/\kappa\} \\
&\quad - \sum_{j=1}^{\tau(S,l,t+1)} c(n_j)\Bigg) \Bigg| \theta^*, B_0 = \mathrm{Beta}(\alpha,\beta), C_0 = C, n_0, X_0\Bigg]\Bigg] \\
&= \mathbb{E}_{\theta^* \sim \mathrm{Beta}(\alpha,\beta)}\mathbb{E}_{n_0 \sim \pi(S,l)}\mathbb{E}_{X_0 \sim \mathrm{Bin}(n_0,\theta^*)}\Bigg[\mathbb{1}\{f(\alpha + X_0, \beta + n_0 + X_0) < 1/\kappa\} \cdot \mathbb{E}_{\pi_l}\Bigg[\\
&\quad \left(\rho^{\mathbb{A}} + \varepsilon \cdot \left(C + c(n_0) + \sum_{j=1}^{\tau(S,l,t+1)} c(n_j)\right)\right) \mathbb{1}\{\exists j \in [\tau(S,l,t+1)] \colon f(\alpha,\beta)\cdot M_{j+1} \geq 1/\kappa\} \\
&\quad - \sum_{j=1}^{\tau(S,l,t+1)} c(n_j)\Bigg] \Bigg| \theta^*, B_0 = \mathrm{Beta}(\alpha,\beta), C_0 = C, n_0, X_0\Bigg]
\end{aligned}
\tag{46}
$$

Here, note that the inner expectation $\mathbb{E}_{\pi_l}[\bullet|\theta^*, B_0 = \mathrm{Beta}(\alpha,\beta), C_0 = C, n_0, X_0]$ is conditioned on n_0 and X_0. Therefore, the first action n_1 is distributed according to $\pi(\alpha + X_0, \beta + n_0 - X_0, C + c(n_0), l+1)$, and any subsequent action n_j is distributed according to:

$$
\begin{cases}
n_1|n_0, X_0 \sim \pi\left(\alpha + X_0, \beta + n_0 - X_0, C + c(n_0), l+1\right) \\
X_1|n_0, X_0 \sim \mathrm{Bin}(n_1, \theta^*) \\
n_2|n_0, X_0 \sim \pi\left(\alpha + X_0 + X_1, \beta + n_0 + n_1 - X_0 - X_1, C + c(n_0) + c(n_1), l+2\right) \\
X_2|n_0, X_0 \sim \mathrm{Bin}(n_2, \theta^*) \\
\quad \vdots \\
n_j|n_0, X_0 \sim \pi\big(\alpha + X_0 + \cdots + X_{j-1}, \beta + n_0 + \cdots + n_{j-1} - X_0 - \cdots - X_{j-1}, \\
\qquad\qquad C + c(n_0) + \cdots + c(n_{j-1}), l+j\big) \\
X_j|n_0, X_0 \sim \mathrm{Bin}(n_j, \theta^*) \\
\quad \vdots
\end{cases}
$$

These are equal in distribution to the sequence of actions and experimental outcomes $\tilde{n}_1, \tilde{X}_1, \ldots$ in the inner expectation of Eq. 45, *i.e.*, we have the following equalities in distribution

$$
\begin{cases}
n_1|n_0, X_0 \overset{d}{=} \tilde{n}_0 \\
X_1|n_0, X_0 \overset{d}{=} \tilde{X}_0 \\
E(n_1, X_1)|n_0, X_0 \overset{d}{=} E(\tilde{n}_0, \tilde{X}_0) \\
\quad \vdots \\
n_j|n_0, X_0 \overset{d}{=} \tilde{n}_{j-1} \\
X_j|n_0, X_0 \overset{d}{=} \tilde{X}_{j-1} \\
E(n_j, X_j)|n_0, X_0 \overset{d}{=} E(\tilde{n}_{j-1}, \tilde{X}_{j-1}) \\
\quad \vdots
\end{cases}
\implies
\begin{cases}
M_1|n_0, X_0 \overset{d}{=} E(n_0, X_0) \cdot \tilde{M}_0 \\
M_2|n_0, X_0 \overset{d}{=} E(n_0, X_0) \cdot \tilde{M}_1 \\
\quad \vdots \\
M_j|n_0, X_0 \overset{d}{=} E(n_0, X_0) \cdot \tilde{M}_{j-1} \\
\quad \vdots
\end{cases}
. \tag{47}
$$

Based on the above, we can also conclude the following about the stopping times under the event $\{f(\alpha, \beta) \cdot E(n_0, X_0) < 1/\kappa\}$:

$$
\begin{aligned}
\tau(S, l, t+1)|n_0, X_0 &\overset{d}{=} (t+1) \wedge \min\{j \in \{1, \ldots, t+1\} \colon n_j = 0 \text{ or } f(\alpha, \beta) \cdot M_{j+1} \geq 1/\kappa\} \\
&\overset{d}{=} (t+1) \wedge \Big(1 + \min\{j \in \{0, \ldots, t\} \colon \tilde{n}_j = 0 \text{ or } \\
&\qquad f(\alpha, \beta) \cdot E(n_0, X_0) \cdot \tilde{M}_{j+1} \geq 1/\kappa\}\Big) \\
&\overset{d}{=} 1 + \underbrace{\tau(S(n_0, X_0), l+1, t)}_{\tilde{\tau}}
\end{aligned} \tag{48}
$$

As a consequence, in Eq. 46, we can change the summation index $j \to j-1$ and use the equalities in distribution in Eq. 47 and Eq. 48 to finally conclude:

$$
\begin{aligned}
\square &= \mathbb{E}_{\theta^* \sim \mathrm{Beta}(\alpha,\beta)} \mathbb{E}_{n_0 \sim \pi(S,l)} \mathbb{E}_{X_0 \sim \mathrm{Bin}(n_0,\theta^*)} \Bigg[\mathbb{1}\{f(\alpha + X_0, \beta + n_0 + X_0) < 1/\kappa\} \cdot \mathbb{E}_{\pi_{l+1}} \Bigg[\\
&\left(\rho^{\mathbb{A}} + \varepsilon \cdot \left(C + c(n_0) + \sum_{j=0}^{\tilde{\tau}} c(\tilde{n}_j) \right) \right) \cdot \mathbb{1}\{\exists j \in [\tilde{\tau}] \colon f(\alpha + X_0, \beta + n_0 - X_0) \cdot \tilde{M}_{j+1} \geq 1/\kappa\} \\
&- \left(\sum_{j=0}^{\tilde{\tau}} c(\tilde{n}_j) \right) \Bigg| \theta^*, B_0 = \mathrm{Beta}(\alpha + X_0, \beta + n_0 - X_0), C_0 = C + c(n_0) \Bigg] \Bigg] \\
&= \ddagger
\end{aligned}
$$

That is, we have shown that $\bar{U}^{\mathbb{A}}_{l,t+1}(\pi|S) = \dagger + \ddagger = V_\pi(S, l, t+1)$. This concludes the induction step and the proof.

F.5. Proof of Proposition 4

Consider the state space $\mathcal{S}$ of the MDP $\mathcal{M}^{\varepsilon}$. We define the set $\mathcal{S}^{\mathrm{r}}$ of *reachable* states from the initial state $(\alpha_0, \beta_0, 0)$ to be any state that can be reached with non-negative probability by a policy π. More formally, $S \in \mathcal{S}^{\mathrm{r}}$ if an only if $S = S^{\mathtt{out}}$ or $S = (\alpha, \beta, C)$ and there exist a $0 \leq t \leq T$ and a sequence of actions and outcomes $n_0, X_0, \dots n_t, X_t$ such that

$$\begin{cases} 0 \leq X_k \leq n_k, \quad \text{for } k = 0, \dots, t \\ \alpha = \alpha_0 + \sum_{k=0}^{t} X_k \\ \beta = \beta_0 + \sum_{k=0}^{t} (n_k - X_k) \\ C = \sum_{k=0}^{t} c(n_t). \end{cases} \tag{49}$$

In words, the sequence $(n_0, X_0, \dots n_t, X_t)$ allows the initial state to eventually transition to the state S according to the transition dynamics (Eq. 8). Observe that then, any state visited under any realization of the MDP for any policy π is contained in the set $\mathcal{S}^{\mathrm{r}}$. Then, $\mathcal{S}^{\mathrm{r}}$ is finite because, for a fixed t, any sequence $(n_0, X_0, \dots n_t, X_t)$ can only take finitely many values (since $|\mathcal{A}| = n^{\mathtt{max}} + 1$), and the MDP has only finitely many steps, namely, $T + 1$.

We focus now on the case where the cost function c is linear. Let $c(n) = c_0 + c_1 \cdot n$, and consider any state that can be reached at time step t with a total number $N = \sum_{k=0}^{t} n_t$, $X = \sum_{k=0}^{t} X_t$. Then,

$$\sum_{k=0}^{t} c(n_t) = (t+1) \cdot c_0 + N \cdot c_1.$$

In particular, in light of Eq. 49, the states in $\mathcal{S}^{\mathrm{r}}$ (except $S^{\mathtt{out}}$) are in a bijection with the triplets (t, N, X). This is because α and β are uniquely determined by X and N, and if $c_0, c_1 \neq 0$, then t and N uniquely determine C.

Next, observe that if the policy never opts out, given a time step $0 \leq t \leq T$, the minimum value that N can take is $N = t+1$, which corresponds to a sequence of actions $n_0 = 1, \dots, n_t = 1$. On the other hand, the maximum value that N can take is $N = (t+1) \cdot n^{\mathtt{max}}$, which corresponds to a sequence of actions $n_0 = n^{\mathtt{max}}, \dots, n_t = n^{\mathtt{max}}$. Further, for a fixed value of N, the total positive outcomes X can take exactly $N + 1$ values, *i.e.*, $X \in \{0, \dots, N\}$. Thus,

$$\begin{aligned} |\mathcal{S}^{\mathrm{r}}| &= \underbrace{1}_{S^{\mathtt{out}}} + \sum_{t=0}^{T} \sum_{N=t+1}^{(t+1)\cdot n^{\mathtt{max}}} (N+1) \\ &= 1 + \sum_{t=0}^{T} \frac{(t+1) \cdot n^{\mathtt{max}} - t}{2} \left((t+2) + ((t+1) \cdot n^{\mathtt{max}} + 1) \right) \\ &= 1 + \sum_{t=0}^{T} \left(\frac{(n^{\mathtt{max}})^2 - 1}{2} \cdot (t+1)^2 + \frac{3n^{\mathtt{max}} - 1}{2} \cdot (t+1) + 1 \right) \\ &= 1 + \frac{(n^{\mathtt{max}})^2 - 1}{12} \cdot (T+1)(T+2)(2T+3) + \frac{3n^{\mathtt{max}} - 1}{4} \cdot (T+1)(T+2) + (T+1) \\ &= \mathcal{O}\left((n^{\mathtt{max}})^2 \cdot T^3 \right) \end{aligned}$$

where we have used standard summation formulas for the arithmetic progression $\sum_{t=0}^{T} (t+1)$ and the quadratic progression $\sum_{t=0}^{T} (t+1)^2$. In particular, the above also shows that the set of reachable states in t time steps, which we denote by $\mathcal{S}^{\mathrm{r}}(t)$, satisfies $|\mathcal{S}^{\mathrm{r}}(t)| = \mathcal{O}\left((n^{\mathtt{max}})^2 \cdot t^2 \right)$.

F.6. Auxiliary Lemma for Proposition 5

Lemma 2. *Consider a state $S \in \mathcal{S}$ such that $S \neq S^{\mathtt{out}}$ and $f(S) < 1/\kappa$, any policy π and $0 \leq l \leq T$. Write $S = (\alpha, \beta, C)$. Then,*

$$\mathbb{E}_\pi \left[\sum_{t=l}^{T} r^\varepsilon(S_t, n_t, S_{t+1}) \middle| S_l = S \right] \leq \rho^A + \varepsilon \cdot C. \tag{50}$$

Note that as an immediate consequence of the above, it also holds for the optimal value function (Eq. 51):

$$V^\varepsilon(S, l) \leq \rho^A + \varepsilon \cdot C$$

Proof. Fix any such S, l and policy π. Denote by τ the corresponding stopping time when $S_l = S$, *i.e.*,

$$\tau = T \wedge \min\{t \in \{l, \dots, T\} \colon S_t = S^{\mathtt{out}} \text{ or } f(S_t) \geq 1/\kappa \mid S_l = S\}.$$

Then, $r^\varepsilon(S_t, n_t, S_{t+1}) = 0$ for $t \geq \tau$ by definition of the rewards (Eq. 9). Thus,

$$\begin{aligned}
\sum_{t=l}^{T} r^\varepsilon(S_t, n_t, S_{t+1}) &= \sum_{t=l}^{\tau-1} r^\varepsilon(S_t, n_t, S_{t+1}) \\
&= \sum_{t=l}^{\tau-1} \left(-c(n_t) + (\rho^{\mathtt{A}} + \varepsilon \cdot (C_t + c(n_t))) \cdot \mathbb{1}\{f(S_{t+1}) \geq 1/\kappa\}\right) \\
&= -\sum_{t=l}^{\tau-1} c(n_t) + \sum_{t=l}^{\tau-1} (\rho^{\mathtt{A}} + \varepsilon \cdot (C_t + c(n_t))) \cdot \mathbb{1}\{f(S_{t+1}) \geq 1/\kappa\}.
\end{aligned}$$

Note that by definition of τ, the second term is at most non-null for the summand with $t = \tau - 1$. Thus:

$$\begin{aligned}
\sum_{t=l}^{T} r^\varepsilon(S_t, n_t, S_{t+1}) &\leq -\sum_{t=l}^{\tau-1} c(n_t) + (\rho^{\mathtt{A}} + \varepsilon \cdot (C_{\tau-1} + c(n_{\tau-1}))) \\
&= -\sum_{t=l}^{\tau-1} c(n_t) + \rho^{\mathtt{A}} + \varepsilon \cdot (C_{\tau-1} + c(n_{\tau-1})) \\
&= -\sum_{t=l}^{\tau-1} c(n_t) + \rho^{\mathtt{A}} + \varepsilon \cdot C_\tau \\
&= \rho^{\mathtt{A}} - \sum_{t=l}^{\tau-1} c(n_t) + \varepsilon \cdot \left(C + \sum_{t=l}^{\tau-1} c(n_t) \right) \\
&\leq \rho^{\mathtt{A}} + \varepsilon \cdot C,
\end{aligned}$$

where in the last step we have used $\varepsilon \leq \varepsilon^{\mathtt{max}} \leq 1$. This concludes the proof.

□

F.7. Proof of Proposition 5

We fix any subsidy $\varepsilon \in [0, \varepsilon^{\texttt{max}}]$. We will use the Bellman optimality equation (Sutton & Barto, 2018), which states that the optimal value function V^ε for $\mathcal{M}^\varepsilon$, defined by:

$$V^\varepsilon(S,l) = \begin{cases} \sup_{\pi\in\Pi} \mathbb{E}_\pi \left[\sum_{t=l}^{T} r^\varepsilon(S_t, n_t, S_{t+1}) \middle| S_l = S \right] & \text{if} \quad 0 \le l \le T \\ 0 & \text{if} \quad l = T+1, \end{cases} \tag{51}$$

satisfies the following recursive condition:

$$V^\varepsilon(S,l) = \max_{n\in\mathcal{A}} \left\{ \sum_{S'\in\mathcal{S}} P(S'|S,n) \cdot (r^\varepsilon(S,n,S') + V^\varepsilon(S',l+1)) \right\},$$

where P are the transition dynamics of the MDP, defined implicitly in Eq. 8. We can now particularize to the action space $\mathcal{A}$ in $\mathcal{M}^\varepsilon$. Since the action $n = 0$ results in a null reward (see Eq. 9), we obtain for $l \le T$ and any state $S = (\alpha, \beta, C)$ such that $f(S) < 1/\kappa$ (and $S \neq S^{\texttt{out}}$):

$$V^\varepsilon(S,l) = \max \left\{ 0, \max_{n\in\{1,\dots,n^{\texttt{max}}\}} \mathbb{E}_{X\sim \mathrm{BB}(n,\alpha,\beta)} \left[r^\varepsilon(S,n,S(n,X)) + V^\varepsilon(S(n,X),l+1) \right] \right\} \tag{52}$$

while $V^\varepsilon(S,l) = 0$ if $f(S) \ge 1/\kappa$ or $S = S^{\texttt{out}}$. In the above, to simplify notation, we denote by $S(n,X)$ the state to which the MDP transitions from S after selecting $n > 0$ and observing outcome X, *i.e.*,

$$S(n,X) = (\alpha + X, \beta + n - X, C + c(n)).$$

F.7.1. Monotonicity on the belief

We focus on proving the monotonicity property in Proposition 5 of V^ε in the parameter α.[18] More precisely, fix a state $S = (\alpha, \beta, C)$ such that $f(\alpha, \beta) < 1/\kappa$ and $\alpha' \ge \alpha$ such that $f(\alpha', \beta) < 1/\kappa$. Denote $S' = (\alpha', \beta, C)$. We want to show that

$$\boxed{V^\varepsilon(\alpha,\beta,C,l) \le V^\varepsilon(\alpha',\beta,C,l)}$$

for any $0 \le l \le T$. We proceed by induction over l.

Base case $l = T$**.** Using Eq. 52 we obtain:

[18] The monotonicity in the parameter β follows using a completely symmetric argument.

$$
\begin{aligned}
&V^{\varepsilon}(\alpha', \beta, C, T) \\
&= \max\left\{0, \max_{n\in\{1,\dots,n^{\max}\}} \mathbb{E}_{X\sim \mathrm{BB}(n,\alpha',\beta)}\left[r^{\varepsilon}(S', n, S'(n,X))\right]\right\} \\
&= \max\Big\{0, \max_{n\in\{1,\dots,n^{\max}\}} \mathbb{E}_{X\sim \mathrm{BB}(n,\alpha',\beta)}\Big[-c(n) \\
&\quad + (\rho^{\mathbb{A}} + \varepsilon\cdot(C + c(n)))\cdot \mathbb{1}\{f(\alpha',\beta)\cdot E(n,X) \geq 1/\kappa\}\Big]\Big\} \\
&\overset{(*)}{\geq} \max\Big\{0, \max_{n\in\{1,\dots,n^{\max}\}} \mathbb{E}_{X\sim \mathrm{BB}(n,\alpha',\beta)}\Big[-c(n) \\
&\quad + (\rho^{\mathbb{A}} + \varepsilon\cdot(C + c(n)))\cdot \mathbb{1}\{f(\alpha,\beta)\cdot E(n,X) \geq 1/\kappa\}\Big]\Big\} \\
&\overset{(**)}{\geq} \max\Big\{0, \max_{n\in\{1,\dots,n^{\max}\}} \mathbb{E}_{X\sim \mathrm{BB}(n,\alpha,\beta)}\Big[-c(n) \\
&\quad + (\rho^{\mathbb{A}} + \varepsilon\cdot(C + c(n)))\cdot \mathbb{1}\{f(\alpha,\beta)\cdot E(n,X) \geq 1/\kappa\}\Big]\Big\} \\
&= V^{\varepsilon}(\alpha, \beta, C, T)
\end{aligned}
$$

where in $(*)$ we have used that the function $f(\bullet,\bullet)$ defined in Eq. 3 is non-decreasing in its first component[19]. In step $(**)$ we have used that the distribution $\mathrm{BB}(\alpha',\beta,n)$ stochastically dominates (in the first-order sense) $\mathrm{BB}(\alpha,\beta,n)$ if $\alpha' \geq \alpha$.

Induction step $l+1 \to l$. Suppose now that $V^{\varepsilon}(\alpha,\beta,C,l+1) \leq V^{\varepsilon}(\alpha',\beta,C,l+1)$ holds for a certain $l+1 \leq T$ and for all α,β,α' such that $\alpha' \geq \alpha$, $f(\alpha,\beta) < 1/\kappa$ and $f(\alpha',\beta) < 1/\kappa$. Using Eq. 52 again, note that:

$$
V^{\varepsilon}(\alpha',\beta,C,l) = \max\left\{0, \max_{n\in\{1,\dots,n^{\max}\}} \underbrace{\mathbb{E}_{X\sim \mathrm{BB}(n,\alpha',\beta)}\left[r^{\varepsilon}(S',n,S'(n,X)) + V^{\varepsilon}(S'(n,X),l+1)\right]}_{\dagger}\right\}
$$

Consider a given $n > 0$ and expand the term $\dagger$ above:

$$
\begin{aligned}
\dagger &= \mathbb{E}_{X\sim \mathrm{BB}(n,\alpha',\beta)}\left[r^{\varepsilon}(S',n,S'(n,X)) + V^{\varepsilon}(S'(n,X),l+1)\right] \\
&= \mathbb{E}_{X\sim \mathrm{BB}(n,\alpha',\beta)}\Big[\mathbb{1}\{f(S'(n,X)) \geq 1/\kappa\}\left(r^{\varepsilon}(S',n,S'(n,X)) + V^{\varepsilon}(S'(n,X),l+1)\right) \\
&+ \mathbb{1}\{f(S'(n,X)) < 1/\kappa\}\left(r^{\varepsilon}(S',n,S'(n,X)) + V^{\varepsilon}(S'(n,X),l+1)\right)\Big] \\
&\overset{(*)}{\geq} \mathbb{E}_{X\sim \mathrm{BB}(n,\alpha',\beta)}\Big[\mathbb{1}\{f(S'(n,X)) \geq 1/\kappa\}\left(r^{\varepsilon}(S',n,S'(n,X)) + V^{\varepsilon}(S'(n,X),l+1)\right) \\
&+ \mathbb{1}\{f(S'(n,X)) < 1/\kappa\}\left(r^{\varepsilon}(S',n,S'(n,X)) + V^{\varepsilon}(S(n,X),l+1)\right)\Big] \\
&\overset{(**)}{=} \mathbb{E}_{X\sim \mathrm{BB}(n,\alpha',\beta)}\Big[\mathbb{1}\{f(S'(n,X)) \geq 1/\kappa, f(S(n,X)) \geq 1/\kappa\} \\
&\quad \cdot \left(r^{\varepsilon}(S',n,S'(n,X)) + V^{\varepsilon}(S(n,X),l+1)\right) \\
&+ \mathbb{1}\{f(S'(n,X)) \geq 1/\kappa, f(S(n,X)) < 1/\kappa\} \\
&\quad \cdot \left(r^{\varepsilon}(S',n,S'(n,X)) + V^{\varepsilon}(S'(n,X),l+1)\right) \\
&+ \mathbb{1}\{f(S'(n,X)) < 1/\kappa\}\left(r^{\varepsilon}(S',n,S'(n,X)) + V^{\varepsilon}(S(n,X),l+1)\right)\Big]
\end{aligned}
$$

[19]This is immediate to verify since $\theta^{\flat} \in (0,1) \implies \log(1+\theta^{\flat}(e-1)) \in (0,1)$.

where in $(*)$ we have used that if $f(S'(n,X)) < 1/\kappa$, then $f(S(n,X)) < 1/\kappa$ and hence the induction hypothesis applies $V^\varepsilon(S'(n,X), l+1) \geq V^\varepsilon(S(n,X), l+1)$. In $(**)$ we have used that if $f(S'(n,X)) \geq 1/\kappa$ and $f(S(n,X)) \geq 1/\kappa$, then $V^\varepsilon(S(n,X), l+1) = V^\varepsilon(S'(n,X), l+1) = 0$.

Lastly, in †, consider the case $f(S'(n,X)) \geq 1/\kappa$ and $f(S(n,X)) < 1/\kappa$. Then,

$$r^\varepsilon(S', n, S'(n,X)) + V^\varepsilon(S'(n,X), l+1) = -c(n) + (\rho^{\mathbb{A}} + \varepsilon \cdot (C + c(n)))$$

and,

$$r^\varepsilon(S, n, S(n,X)) + V^\varepsilon(S(n,X), l+1) = -c(n) + V^\varepsilon(S(n,X), l+1) \overset{\diamond}{\leq} -c(n) + \rho^{\mathbb{A}} + \varepsilon \cdot (C + c(n))),$$

where $\diamond$ follows from Lemma 2.

Finally, noting that $r^\varepsilon(S', n, S'(n,X)) \geq r^\varepsilon(S, n, S(n,X))$ and using the first-order stochastic dominance for the Beta-Binomial again, we conclude that:

$$\begin{aligned}
\dagger \geq \mathbb{E}_{X\sim \mathrm{BB}(n,\alpha',\beta)}\Big[&\mathbb{1}\{f(S'(n,X)) \geq 1/\kappa,\ f(S(n,X)) \geq 1/\kappa\} \\
&\cdot \big(r^\varepsilon(S, n, S(n,X)) + V^\varepsilon(S(n,X), l+1)\big) \\
&+ \mathbb{1}\{f(S'(n,X)) \geq 1/\kappa,\ f(S(n,X)) < 1/\kappa\} \\
&\cdot \big(r^\varepsilon(S, n, S(n,X)) + V^\varepsilon(S(n,X), l+1)\big) \\
&+ \mathbb{1}\{f(S'(n,X)) < 1/\kappa\} \\
&\cdot \big(r^\varepsilon(S, n, S(n,X)) + V^\varepsilon(S(n,X), l+1)\big)\Big] \\
\geq \mathbb{E}_{X\sim \mathrm{BB}(n,\alpha,\beta)}&\big[r^\varepsilon(S, n, S(n,X)) + V^\varepsilon(S(n,X), l+1)\big].
\end{aligned}$$

That is,

$$\begin{aligned}
\mathbb{E}_{X\sim \mathrm{BB}(n,\alpha',\beta)}&\left[r^\varepsilon(S', n, S'(n,X)) + V^\varepsilon(S'(n,X), l+1)\right] \geq \\
&\mathbb{E}_{X\sim \mathrm{BB}(n,\alpha,\beta)}\left[r^\varepsilon(S, n, S(n,X)) + V^\varepsilon(S(n,X), l+1)\right], \quad (53)
\end{aligned}$$

and taking maximum over the action n and using Eq. 52:

$$V^\varepsilon(\alpha', \beta, C, l) \geq V^\varepsilon(\alpha, \beta, C, l).$$

This concludes the induction step and thus the proof of the monotonicity in the belief.

F.7.2. Monotonicity on the cost

We now focus on proving the monotonicity property in Proposition 5 of V^ε in the cumulated cost C. More precisely, fix a state $S = (\alpha, \beta, C)$ such that $f(\alpha, \beta) < 1/\kappa$ and consider $C' \geq C$. Denote $S' = (\alpha, \beta, C')$. We want to show that

$$\boxed{V^\varepsilon(\alpha, \beta, C, l) \leq V^\varepsilon(\alpha, \beta, C', l)}$$

for any $0 \leq l \leq T$. Again, we proceed by induction over l.

Base $l = T$**.** Using Eq. 52 we obtain:

$$
\begin{aligned}
V^{\varepsilon}(\alpha,\beta,C',T) &= \max\Big\{0, \max_{n\in\{1,\dots,n^{\max}\}} \mathbb{E}_{X\sim \mathrm{BB}(n,\alpha,\beta)}[r^{\varepsilon}(S',n,S'(n,X))]\Big\} \\
&= \max\Big\{0, \max_{n\in\{1,\dots,n^{\max}\}} \mathbb{E}_{X\sim \mathrm{BB}(n,\alpha,\beta)}\Big[-c(n) \\
&\qquad + \big(\rho^{\mathbb{A}} + \varepsilon(C'+c(n))\big)\cdot \mathbb{1}\{f(\alpha,\beta)\cdot E(n,X)\geq 1/\kappa\}\Big]\Big\} \\
&\geq \max\Big\{0, \max_{n\in\{1,\dots,n^{\max}\}} \mathbb{E}_{X\sim \mathrm{BB}(n,\alpha,\beta)}\Big[-c(n) \\
&\qquad + \big(\rho^{\mathbb{A}} + \varepsilon(C+c(n))\big)\cdot \mathbb{1}\{f(\alpha,\beta)\cdot E(n,X)\geq 1/\kappa\}\Big]\Big\} \\
&= V^{\varepsilon}(\alpha,\beta,C,T).
\end{aligned}
$$

Induction step $l+1\to l$**.** Suppose now that $V^{\varepsilon}(\alpha,\beta,C,l+1)\leq V^{\varepsilon}(\alpha,\beta,C',l+1)$ holds for a certain $l+1\leq T$ and for all α,β such that $f(\alpha,\beta)<1/\kappa$ and $0\leq C\leq C'$. Using Eq. 52 again, note that:

$$
V^{\varepsilon}(\alpha,\beta,C',l) = \max\left\{0, \max_{n\in\{1,\dots,n^{\max}\}} \underbrace{\mathbb{E}_{X\sim \mathrm{BB}(n,\alpha,\beta)}\left[r^{\varepsilon}(S',n,S'(n,X)) + V^{\varepsilon}(S'(n,X),l+1)\right]}_{\dagger}\right\}
$$

Consider a given $n>0$ and expand the term † above:

$$
\begin{aligned}
\dagger &= \mathbb{E}_{X\sim \mathrm{BB}(n,\alpha,\beta)}\left[r^{\varepsilon}(S',n,S'(n,X)) + V^{\varepsilon}(S'(n,X),l+1)\right] \\
&= \mathbb{E}_{X\sim \mathrm{BB}(n,\alpha,\beta)}\Big[\mathbb{1}\{f(S'(n,X))\geq 1/\kappa\}\left(r^{\varepsilon}(S',n,S'(n,X)) + V^{\varepsilon}(S'(n,X),l+1)\right) \\
&\quad + \mathbb{1}\{f(S'(n,X))< 1/\kappa\}\left(r^{\varepsilon}(S',n,S'(n,X)) + V^{\varepsilon}(S'(n,X),l+1)\right)\Big] \\
&\overset{(*)}{\geq} \mathbb{E}_{X\sim \mathrm{BB}(n,\alpha,\beta)}\Big[\mathbb{1}\{f(S'(n,X))\geq 1/\kappa\}\left(r^{\varepsilon}(S',n,S'(n,X)) + V^{\varepsilon}(S'(n,X),l+1)\right) \\
&\quad + \mathbb{1}\{f(S'(n,X))< 1/\kappa\}\left(r^{\varepsilon}(S',n,S'(n,X)) + V^{\varepsilon}(S(n,X),l+1)\right)\Big] \\
&\overset{(**)}{=} \mathbb{E}_{X\sim \mathrm{BB}(n,\alpha,\beta)}\Big[\mathbb{1}\{f(S'(n,X))\geq 1/\kappa\}\left(r^{\varepsilon}(S',n,S'(n,X)) + V^{\varepsilon}(S(n,X),l+1)\right) \\
&\quad + \mathbb{1}\{f(S'(n,X))< 1/\kappa\}\left(r^{\varepsilon}(S',n,S'(n,X)) + V^{\varepsilon}(S(n,X),l+1)\right)\Big]
\end{aligned}
$$

where in $(*)$ we have used that if $f(S'(n,X))<1/\kappa$, then the induction hypothesis applies $V^{\varepsilon}(S'(n,X),l+1)\geq V^{\varepsilon}(S(n,X),l+1)$ because the total cost in state $S'(n,X)$ is $C'+c(n)\geq C+c(n)$, which equals the total cost in $S(n,X)$. In $(**)$ we have used that if $f(S'(n,X))\geq 1/\kappa$, then $f(S(n,X))\geq 1/\kappa$ and hence $V^{\varepsilon}(S'(n,X),l+1) = V^{\varepsilon}(S(n,X),l+1)=0$.

Finally, noting that $r^{\varepsilon}(S',n,S'(n,X))\geq r^{\varepsilon}(S,n,S(n,X))$, we conclude that:

$$
\begin{aligned}
\dagger &\geq \mathbb{E}_{X\sim \mathrm{BB}(n,\alpha,\beta)}\Big[\mathbb{1}\{f(S'(n,X))\geq 1/\kappa\}\left(r^{\varepsilon}(S,n,S(n,X)) + V^{\varepsilon}(S(n,X),l+1)\right) \\
&\quad + \mathbb{1}\{f(S'(n,X))< 1/\kappa\}\left(r^{\varepsilon}(S,n,S(n,X)) + V^{\varepsilon}(S(n,X),l+1)\right)\Big] \\
&= \mathbb{E}_{X\sim \mathrm{BB}(n,\alpha,\beta)}\left[r^{\varepsilon}(S,n,S(n,X)) + V^{\varepsilon}(S(n,X),l+1)\right]
\end{aligned}
$$

That is,

$$\mathbb{E}_{X\sim\mathrm{BB}(n,\alpha,\beta)}\left[r^{\varepsilon}(S',n,S'(n,X))+V^{\varepsilon}(S'(n,X),l+1)\right]\geq \\ \mathbb{E}_{X\sim\mathrm{BB}(n,\alpha,\beta)}\left[r^{\varepsilon}(S,n,S(n,X))+V^{\varepsilon}(S(n,X),l+1)\right],$$

and taking maximum over the action n and using Eq. 52:

$$V^{\varepsilon}(\alpha,\beta,C,l)\geq V^{\varepsilon}(\alpha,\beta,C',l).$$

This concludes the induction step and thus the proof of the monotonicity in the cost.

F.8. Proof of Proposition 6

Fix any subsidy level $\varepsilon \in [0, \varepsilon^{\texttt{max}}]$, total cost $C \geq 0$, and an initial time step $0 \leq l \leq T$. Consider any state $S \neq S^{\texttt{out}}$ and write $S = (\alpha, \beta, C)$. Then, the Bellman optimality condition (Sutton & Barto, 2018) establishes that the action $n = 0$ is optimal if and only if the value function coincides with the value of action $n = 0$ (which is 0 by definition of the reward function in Eq. 9), that is,

$$V^{\varepsilon}(S, l) = 0,$$

where $V^{\varepsilon}(S, l)$ is the optimal value function defined in Eq. 51. Using Eq. 52, this is equivalent to the condition:

$$\begin{aligned} &\max\left\{0, \underbrace{\max_{n \in \{1,\ldots,n^{\texttt{max}}\}} \mathbb{E}_{X \sim \mathrm{BB}(n,\alpha,\beta)}\left[r^{\varepsilon}(S, n, S(n, X)) + V^{\varepsilon}(S(n, X), l+1)\right]}_{=H(\alpha,\beta,C)}\right\} = 0 \\ \iff &H(\alpha, \beta, C) \leq 0. \end{aligned}$$

Now, we observe that in proving Proposition 5 in Appendix F.7 (see Eq. 53), we precisely showed by induction that the function $H(\alpha, \beta, C)$ is non-decreasing in α, and an analogous argument shows that it is non-increasing in β. We leverage this property in what follows.

For any $\alpha > 0$, define the quantity

$$\tilde{\beta}(\alpha) = \sup\{\beta : \ H(\alpha, \beta, C) > 0 \text{ and } \beta > 0\}.$$

As a consequence of the monotonicity of $H(\alpha, \beta, C)$, if $\alpha' \geq \alpha$, then

$$H(\alpha, \beta, C) > 0 \implies H(\alpha', \beta, C) > 0,$$

and thus the following set inclusion holds:

$$\{\beta : \ H(\alpha, \beta, C) > 0 \text{ and } \beta > 0\} \subseteq \{\beta : \ H(\alpha', \beta, C) > 0 \text{ and } \beta > 0\}.$$

Taking the supremum, the above implies:

$$\underbrace{\sup\{\beta : \ H(\alpha, \beta, C) > 0 \text{ and } \beta > 0\}}_{\tilde{\beta}(\alpha)} \leq \underbrace{\sup\{\beta : \ H(\alpha', \beta, C) > 0 \text{ and } \beta > 0\}}_{\tilde{\beta}(\alpha')},$$

which shows that $\tilde{\beta}(\alpha)$ is non-decreasing in α.

Next, observe that:

- If $\beta > \tilde{\beta}(\alpha)$, then by definition of the supremum we have that $H(\alpha, \beta, C) \leq 0$, and thus the action $n = 0$ is optimal.
- If $\beta < \tilde{\beta}(\alpha)$, then again by the definition of the supremum, there exists β' such that $\beta < \beta' \leq \tilde{\beta}(\alpha)$ with $H(\alpha, \beta', C) > 0$. Then, using the monotonicity, $H(\alpha, \beta, C) \geq H(\alpha, \beta', C) > 0$, which means that opting out is strictly sub-optimal: there exists an action $n > 0$ that leads to a strictly higher expected value if taken at state S and time l.

This proves the claim in Proposition 6.

F.9. Proof of Proposition 7

Recall that the value function in the MDP $\mathcal{M}^\varepsilon$ is defined by:

$$V_\pi^\varepsilon(S,l) = \mathbb{E}_\pi\left[\sum_{t=l}^{T} r^\varepsilon(S_t, n_t, S_{t+1}) \middle| S_l = S\right], \tag{54}$$

with r^ε defined by:

$$r^\varepsilon(S,n,S') = \begin{cases} -c(n) + (\rho^{\mathrm{A}} + \varepsilon\cdot(C + c(n)))\cdot \mathbb{1}\{f(S') \geq 1/\kappa\} & \text{if } 0 < f(S) < 1/\kappa \\ 0 & \text{if } S = S^{\mathtt{out}} \text{ or } f(S) \geq 1/\kappa. \end{cases} \tag{55}$$

Firstly, note that all rewards in Eq. 54 become 0 as soon as $S_t = S^{\mathtt{out}}$ or $f(S_t) \geq 1/\kappa$. This is because, from the transition dynamics in Eq. 8, any such state S_t is absorbing, *i.e.*, $S_{t+1} = S_t$ if $S_t = S^{\mathtt{out}}$ or $f(S_t) \geq 1/\kappa$, and satisfies $r^\varepsilon(S_t, n_t, S_{t+1}) = 0$. Then, the terms in Eq. 54 that are non-zero correspond to time steps $t \leq \tau$, where τ is the stopping time defined by:

$$\tau = T \wedge \min\{t \in \{l,\dots,T\} \colon n_t = 0 \,\text{or}\, f(S_{t+1}) \geq 1/\kappa\}.$$

For $l \leq t \leq \tau$, the first case of the reward function in Eq. 55 applies:

$$\begin{aligned} V_\pi^\varepsilon(S,l) &= \mathbb{E}_\pi\left[\sum_{t=l}^{\tau} (-c(n_t) + (\rho^{\mathrm{A}} + \varepsilon C_{t+1})\mathbb{1}\{f(S_{t+1}) \geq 1/\kappa\}) \,\middle|\, S_l = S\right] \\ &= \mathbb{E}_\pi\left[\sum_{t=l}^{\tau} \left(-c(n_t) + \rho^{\mathrm{A}}\mathbb{1}\{f(S_{t+1}) \geq 1/\kappa\}\right) \right. \\ &\qquad \left. + \varepsilon \sum_{t=l}^{\tau} C_{t+1}\mathbb{1}\{f(S_{t+1}) \geq 1/\kappa\} \,\middle|\, S_l = S\right] \end{aligned}$$

Note that the first part of the expectation corresponds exactly to the value function under no subsidy, $V_\pi^0(S,l)$, and thus,

$$V_\pi^\varepsilon(S,l) = V_\pi^0(S,l) + \varepsilon\cdot\mathbb{E}_\pi\left[\sum_{t=l}^{\tau} C_{t+1}\cdot\mathbb{1}\{f(S_{t+1}) \geq 1/\kappa\} \,\middle|\, S_l = S\right]$$

Now, define:

$$\begin{aligned} A_\pi(S,l) &= \mathbb{E}_\pi\left[\sum_{t=l}^{\tau} C_{t+1}\cdot\mathbb{1}\{f(S_{t+1}) \geq 1/\kappa\} \,\middle|\, S_l = S\right] \\ &= \mathbb{E}_\pi\left[C_{\tau+1}\cdot\mathbb{1}\{f(S_{\tau+1}) \geq 1/\kappa\} \mid S_l = S\right] \end{aligned} \tag{56}$$

This term corresponds to the (expected) total cost incurred by the agent conditional on the product being approved, when the MDP starts from state S at time l.[20] Indeed, by the definition of the stopping time τ, approval (*i.e.*, $f(S_{t+1}) \geq 1/\kappa$) can only happen at exactly $t = \tau$. If the agent opts out or the horizon T is reached without approval, the indicator $\mathbb{1}\{f(S_{t+1}) \geq 1/\kappa\}$ is 0 for all t. We conclude that

$$V_\pi^\varepsilon(S,l) = V_\pi^0(S,l) + \varepsilon\cdot A_\pi(S,l).$$

[20] We use the term "conditional on approval" informally and for didactic purposes, since $A_\pi(S,l)$ is the quantity that naturally appears in the value decomposition $V_\pi^\varepsilon(S,l) = V_\pi^0(S,l) + \varepsilon\cdot A_\pi(S,l)$, representing the expected subsidy paid by the principal. Strictly speaking, however, $A_\pi(S,l)$ is the expected cost weighted by the indicator of approval, rather than a conditional expectation in the measure-theoretic sense.

F.10. Proof of Proposition 8

Denote by $\Pi^{\mathrm{r}} \subset \Pi$ the set of all deterministic policies that select action $n = 0$ at $S^{\mathtt{out}}$, any state such that $f(S) \geq 1/\kappa$, or any state such that $S \notin S^{\mathrm{r}}$, and observe that Π^{r} is finite. Recall from Proposition 7 that, for any policy $\pi \in \Pi^{\mathrm{r}}$,

$$V_\pi^\varepsilon(S,l) = V_\pi^0(S,l) + \varepsilon \cdot A_\pi(S,l). \tag{57}$$

In particular, at the initial state $S_0 = (\alpha_0, \beta_0, 0)$ and time $l = 0$, we have

$$\bar{U}^{\mathrm{A}}(\pi;\varepsilon) = V_\pi^\varepsilon(S_0,0) = V_\pi^0(S_0,0) + \varepsilon \cdot A_\pi(S_0,0). \tag{58}$$

Since Π^{r} is finite, for every $(S,l) \in \mathcal{S}^{\mathrm{r}} \times [T]$ the optimal value function

$$V^\varepsilon(S,l) = \max_{\pi \in \Pi^{\mathrm{r}}} \left\{ V_\pi^0(S,l) + \varepsilon \cdot A_\pi(S,l) \right\} \tag{59}$$

is the point-wise maximum of finitely many affine functions in ε. It is well-known that such a point-wise maximum is convex, continuous, and piecewise linear (Rockafellar, 1970). Here, we particularize to our problem, with the goal of constructing a single partition of $[0, \varepsilon^{\mathtt{max}}]$ on which a single optimal policy (for $\mathcal{M}^\varepsilon$ at every state $S \in \mathcal{S}^{\mathrm{r}}$ and time $l \in [T]$ simultaneously) is optimal in each interval of the partition.

For any two distinct policies $\pi, \pi' \in \Pi^{\mathrm{r}}$ and any $(S,l) \in \mathcal{S}^{\mathrm{r}} \times [T]$, consider the difference

$$(V_\pi^0(S,l) - V_{\pi'}^0(S,l)) + \varepsilon(A_\pi(S,l) - A_{\pi'}(S,l)).$$

If $(V_\pi^0(S,l), A_\pi(S,l)) = (V_{\pi'}^0(S,l), A_{\pi'}(S,l))$ for all (S,l), then the two policies yield identical values for all ε and all (S,l), and we may break ties arbitrarily and retain only one of them. Similarly, if $V_\pi^0(S,l) + \varepsilon \cdot A_\pi(S,l) \geq V_{\pi'}^0(S,l) + \varepsilon \cdot A_{\pi'}(S,l)$ for all $\varepsilon \in [0, \varepsilon^{\mathtt{max}}]$ and all (S,l), then policy π' can be removed without changing the optimal value function at any (S,l). Thus, without loss of generality, we restrict our attention to a subset $\tilde{\Pi}^{\mathrm{r}} \subset \Pi^{\mathrm{r}}$ such that for any distinct $\pi, \pi' \in \tilde{\Pi}^{\mathrm{r}}$ and any (S,l), the corresponding affine functions intersect exactly once in $[0, \varepsilon^{\mathtt{max}}]$, and each policy is optimal for some value of ε.

Then, let

$$\mathcal{K} = \big\{ \varepsilon \in [0, \varepsilon^{\mathtt{max}}] : \exists\, \pi \neq \pi' \in \tilde{\Pi}^{\mathrm{r}},\, \exists\, (S,l) \in \mathcal{S}^{\mathrm{r}} \times [T] \\ \text{such that } V_\pi^0(S,l) + \varepsilon \cdot A_\pi(S,l) = V_{\pi'}^0(S,l) + \varepsilon \cdot A_{\pi'}(S,l) \big\}.$$

Since $\tilde{\Pi}^{\mathrm{r}}$, $\mathcal{S}^{\mathrm{r}}$, and $[T]$ are all finite, and each quadruple (π, π', S, l) contributes at most one point to $\mathcal{K}$, the set $\mathcal{K}$ is finite. Ordering its elements and adding the endpoints if necessary, we obtain a partition

$$0 = \varepsilon_0 < \varepsilon_1 < \cdots < \varepsilon_L = \varepsilon^{\mathtt{max}}.$$

By construction, no two affine functions $\varepsilon \mapsto V_\pi^0(S,l) + \varepsilon \cdot A_\pi(S,l)$ intersect in any open interval $(\varepsilon_i, \varepsilon_{i+1})$ at any (S,l). Hence, for every (S,l), the ordering of $\{V_\pi^0(S,l) + \varepsilon \cdot A_\pi(S,l)\}_{\pi \in \tilde{\Pi}^{\mathrm{r}}}$ is constant on each such interval. It follows that there exists a single policy $\pi_i \in \tilde{\Pi}^{\mathrm{r}}$ that is optimal for $\mathcal{M}^\varepsilon$ at every (S,l) simultaneously, and such that

$$V^\varepsilon(S,l) = V_{\pi_i}^0(S,l) + \varepsilon \cdot A_{\pi_i}(S,l) \quad \text{for all } \varepsilon \in [\varepsilon_i, \varepsilon_{i+1}) \text{ and all } (S,l).$$

In particular, specializing to $(S,l) = (S_0, 0)$,

$$\bar{U}^{\mathrm{A}}(\pi^\varepsilon;\varepsilon) = V_{\pi_i}^0(S_0,0) + \varepsilon \cdot A_{\pi_i}(S_0,0) \quad \text{for all } \varepsilon \in [\varepsilon_i, \varepsilon_{i+1}).$$

This establishes the result for $\bar{U}^{\mathrm{A}}(\pi^\varepsilon;\varepsilon)$. Note that $\mathbb{E}_{(\alpha_0,\beta_0)\sim Q}[\bar{U}^{\mathrm{A}}(\pi^\varepsilon;\varepsilon)]$ is then also convex and continuous because it is an average of convex continuous functions that are uniformly bounded (observe that by Lemma 2, we have $0 \leq \bar{U}^{\mathrm{A}}(\pi^\varepsilon;\varepsilon) \leq \rho^{\mathrm{A}}$ uniformly for any initial belief parameters (α_0, β_0), where the lower bound follows from the fact that the agent can always opt out at no cost); continuity of the expectation then follows from the dominated convergence theorem. Lastly, in each interval $[\varepsilon_i, \varepsilon_{i+1})$, $\bar{U}^{\mathrm{A}}(\pi^\varepsilon;\varepsilon)$ is linear, which implies that $\mathbb{E}_{(\alpha_0,\beta_0)\sim Q}[\bar{U}^{\mathrm{A}}(\pi^\varepsilon;\varepsilon)]$ is also linear:

$$\mathbb{E}_{(\alpha_0,\beta_0)\sim Q}[\bar{U}^{\mathrm{A}}(\pi^\varepsilon;\varepsilon)] = \mathbb{E}_{(\alpha_0,\beta_0)\sim Q}[V_{\pi_i}^0(S_0,0)] + \varepsilon \cdot \mathbb{E}_{(\alpha_0,\beta_0)\sim Q}[A_{\pi_i}(S_0,0)] \quad \text{for all } \varepsilon \in [\varepsilon_i, \varepsilon_{i+1}).$$

This concludes the proof.

F.11. Proof of Proposition 9

Recall that the social utility $\bar{U}^{\mathrm{S}}(\varepsilon;\pi)$ is defined as (Eq. 13):

$$\bar{U}^{\mathrm{S}}(\varepsilon;\pi) = \mathbb{E}_{\pi,(\alpha_0,\beta_0)\sim Q}\left[\sum_{t=0}^{T}\left(\rho^{\mathrm{S}} - \varepsilon\cdot C_{t+1}\right)\cdot \mathbb{1}\left\{0 < f(S_t) < 1/\kappa \leq f(S_{t+1})\right\} \;\middle|\; S_0 = (\alpha_0,\beta_0,0)\right].$$

Let τ be the stopping time defined by:

$$\tau = T \wedge \min\{t \in \{0,\dots,T\}\colon n_t = 0 \text{ or } f(S_{t+1}) \geq 1/\kappa\},$$

that is, the last time step before reaching an absorbing state—either $S^{\mathtt{out}}$, for which $f(S^{\mathtt{out}}) = 0$, or any state S such that $f(S) \geq 1/\kappa$. Then, using the linearity of the expectation:

$$\begin{aligned}
&\bar{U}^{\mathrm{S}}(\varepsilon;\pi)\\
&= \mathbb{E}_{\pi,(\alpha_0,\beta_0)\sim Q}\left[\sum_{t=0}^{T}\left(\rho^{\mathrm{S}} - \varepsilon\cdot C_{t+1}\right)\cdot \mathbb{1}\left\{0 < f(S_t) < 1/\kappa \leq f(S_{t+1})\right\} \;\middle|\; S_0 = (\alpha_0,\beta_0,0)\right]\\
&= \rho^{\mathrm{S}}\cdot \mathbb{E}_{\pi,(\alpha_0,\beta_0)\sim Q}\left[\sum_{t=0}^{T}\mathbb{1}\left\{0 < f(S_t) < 1/\kappa \leq f(S_{t+1})\right\} \;\middle|\; S_0 = (\alpha_0,\beta_0,0)\right]\\
&\quad - \varepsilon\cdot \mathbb{E}_{\pi,(\alpha_0,\beta_0)\sim Q}\left[\sum_{t=0}^{\tau}C_{t+1}\cdot \mathbb{1}\left\{0 < f(S_t) < 1/\kappa \leq f(S_{t+1})\right\} \;\middle|\; S_0 = (\alpha_0,\beta_0,0)\right]\\
&= \rho^{\mathrm{S}}\cdot \mathbb{E}_{(\alpha_0,\beta_0)\sim Q}\mathbb{E}_{\pi}\left[\mathbb{1}\left\{\exists t \in [T]\colon 0 < f(S_t) < 1/\kappa \leq f(S_{t+1})\right\} \mid S_0 = (\alpha_0,\beta_0,0)\right]\\
&\quad - \varepsilon\cdot \mathbb{E}_{(\alpha_0,\beta_0)\sim Q}\mathbb{E}_{\pi}\left[C_{\tau+1}\cdot \mathbb{1}\left\{1/\kappa \leq f(S_{\tau+1})\right\} \mid S_0 = (\alpha_0,\beta_0,0)\right]\\
&= \rho^{\mathrm{S}}\cdot \mathbb{E}_{(\alpha_0,\beta_0)\sim Q}\left[P_{\pi}\left(\exists t \in [T]\colon 0 < f(S_t) < 1/\kappa \leq f(S_{t+1}) \mid S_0 = (\alpha_0,\beta_0,0)\right)\right]\\
&\quad - \varepsilon\cdot \mathbb{E}_{(\alpha_0,\beta_0)\sim Q}\mathbb{E}_{\pi}\left[C_{\tau+1}\cdot \mathbb{1}\left\{1/\kappa \leq f(S_{\tau+1})\right\} \mid S_0 = (\alpha_0,\beta_0,0)\right]\\
&= \rho^{\mathrm{S}}\cdot \mathbb{E}_{(\alpha_0,\beta_0)\sim Q}\left[P_{\pi}\left(\exists t \in [T]\colon 1/\kappa \leq f(S_{t+1}) \mid S_0 = (\alpha_0,\beta_0,0)\right)\right]\\
&\quad - \varepsilon\cdot \mathbb{E}_{(\alpha_0,\beta_0)\sim Q}\left[A_{\pi}(\alpha_0,\beta_0,0,0)\right]
\end{aligned}$$

where we have used the definition of A_π in Eq. 56, and the fact that the condition $0 < f(S_t) < 1/\kappa \leq f(S_{t+1})$ can occur at most once at time step $t = \tau$.

F.12. Proof of Proposition 10

In this section, we will show that Algorithm 1 recovers (in a finite number of steps) the partition of the interval $[0, \varepsilon^{\texttt{max}}]$ given by Proposition 8, which we denote by $\mathcal{P}$:[21]

$$\mathcal{P} = \{0 = \varepsilon_0 < \varepsilon_1 < \cdots < \varepsilon_L = \varepsilon^{\texttt{max}}\},$$

where for each interval $[\varepsilon_i, \varepsilon_{i+1})$, there exists a (deterministic) policy π_i that is optimal for any $\varepsilon \in [\varepsilon_i, \varepsilon_{i+1})$, that is (writing $S_0 = (\alpha_0, \beta_0, 0)$),

$$\begin{cases} \bar{U}^{\mathrm{A}}(\pi^\varepsilon; \varepsilon) = V^0_{\pi_i}(S_0, 0) + \varepsilon \cdot A_{\pi_i}(S_0, 0) & \forall \varepsilon \in [\varepsilon_i, \varepsilon_{i+1}) \\ \mathbb{E}_{(\alpha_0,\beta_0)\sim Q}[\bar{U}^{\mathrm{A}}(\pi^\varepsilon; \varepsilon)] = \mathbb{E}_{(\alpha_0,\beta_0)\sim Q}[V^0_{\pi_i}(S_0, 0)] + \varepsilon \cdot \mathbb{E}_{(\alpha_0,\beta_0)\sim Q}[A_{\pi_i}(S_0, 0)] & \forall \varepsilon \in [\varepsilon_i, \varepsilon_{i+1}). \end{cases}$$

Since the above partition does not depend on the belief Q of the principal (Proposition 8), we assume without loss of generality that the principal knows the agent's initial belief $\mathrm{Beta}(\alpha_0, \beta_0)$. Consequently, the expectations $\mathbb{E}_{(\alpha_0,\beta_0)\sim Q}[\bullet]$ simply correspond to evaluating the integrand at the true belief parameters (α_0, β_0). We begin by proving the following lemma.[22]

Lemma 3. *Let $\varepsilon_L < \varepsilon_R$ be two subsidy levels with, respectively, optimal policies π_L and π_R, and value functions (evaluated at the initial state $(\alpha_0, \beta_0, 0)$ and initial time step $l = 0$, which we omit for notational convenience) $V^0_L + \varepsilon \cdot A_L$ and $V^0_R + \varepsilon \cdot A_R$ (see Proposition 7). The following holds:*

1. *If $A_L = A_R$, then $\bar{U}^{\mathrm{A}}(\pi^\varepsilon; \varepsilon) = V^0_L + \varepsilon \cdot A_L$ for all $\varepsilon \in [\varepsilon_L, \varepsilon_R]$.*
2. *If $A_L < A_R$, let $\varepsilon_{int} = \frac{V^0_L - V^0_R}{A_R - A_L}$. Then, $\bar{U}^{\mathrm{A}}(\pi^\varepsilon; \varepsilon) = \max(V^0_L + \varepsilon \cdot A_L, V^0_R + \varepsilon \cdot A_R)$ for all $\varepsilon \in [\varepsilon_L, \varepsilon_R]$ if and only if $\bar{U}^{\mathrm{A}}(\pi^{\varepsilon_{int}}; \varepsilon_{int}) = V^0_L + \varepsilon_{int} \cdot A_L$.*

Proof. For part 1., if $A_L = A_R$, then since $\bar{U}^{\mathrm{A}}(\pi^\varepsilon; \varepsilon)$ is convex, its subgradient must be non-decreasing in ε. Thus, for any $\varepsilon \in (\varepsilon_L, \varepsilon_R)$, we must have $A_L \leq A_{\pi^\varepsilon} \leq A_R$, which implies $A_{\pi^\varepsilon} = A_L$. By continuity and the fact that $\bar{U}^{\mathrm{A}}(\pi^\varepsilon; \varepsilon)$ is the point-wise maximum of affine functions, it follows that $V^0_{\pi^\varepsilon} = V^0_L = V^0_R$, and the value function is a single affine segment on this interval.

For part 2., the direct implication follows because if $\bar{U}^{\mathrm{A}}(\pi^\varepsilon; \varepsilon) = \max(V^0_L + \varepsilon \cdot A_L, V^0_R + \varepsilon \cdot A_R)$ for all $\varepsilon \in [\varepsilon_L, \varepsilon_R]$, then evaluating at ε_{int}, and since $V^0_L + \varepsilon_{int} \cdot A_L = V^0_R + \varepsilon_{int} \cdot A_R$ by definition, we obtain $\bar{U}^{\mathrm{A}}(\pi^{\varepsilon_{int}}; \varepsilon_{int}) = V^0_L + \varepsilon_{int} \cdot A_L$. For the backward direction, we argue by contradiction and suppose $\bar{U}^{\mathrm{A}}(\pi^{\varepsilon_{int}}; \varepsilon_{int}) = V^0_L + \varepsilon_{int} \cdot A_L$ but there exists some $\varepsilon' \in (\varepsilon_L, \varepsilon_R)$ and a policy π' such that $V^0_{\pi'} + \varepsilon' \cdot A_{\pi'} > \max(V^0_L + \varepsilon' \cdot A_L, V^0_R + \varepsilon' \cdot A_R)$. Assume without loss of generality $\varepsilon' \leq \varepsilon_{int}$ (the argument for the case $\varepsilon' \geq \varepsilon_{int}$ is symmetric, with π_R in place of π_L). Since π_L is optimal at ε_L, we have $V^0_L + \varepsilon_L \cdot A_L \geq V^0_{\pi'} + \varepsilon_L \cdot A_{\pi'}$. Combined with the assumption $V^0_{\pi'} + \varepsilon' \cdot A_{\pi'} > V^0_L + \varepsilon' \cdot A_L$ and subtracting, we obtain $(\varepsilon' - \varepsilon_L)(A_{\pi'} - A_L) > 0$, which implies $A_{\pi'} > A_L$ since $\varepsilon' > \varepsilon_L$. Then, at $\varepsilon_{int} \geq \varepsilon'$,

$$\begin{aligned} V^0_{\pi'} + \varepsilon_{int} \cdot A_{\pi'} &= (V^0_{\pi'} + \varepsilon' \cdot A_{\pi'}) + (\varepsilon_{int} - \varepsilon') \cdot A_{\pi'} \\ &> (V^0_L + \varepsilon' \cdot A_L) + (\varepsilon_{int} - \varepsilon') \cdot A_L \\ &= V^0_L + \varepsilon_{int} \cdot A_L, \end{aligned}$$

where the strict inequality uses $V^0_{\pi'} + \varepsilon' \cdot A_{\pi'} > V^0_L + \varepsilon' \cdot A_L$ together with $A_{\pi'} > A_L$ and $\varepsilon_{int} \geq \varepsilon'$. This contradicts the assumption that $\bar{U}^{\mathrm{A}}(\pi^{\varepsilon_{int}}; \varepsilon_{int}) = V^0_L + \varepsilon_{int} \cdot A_L$. □

To prove Proposition 10, we begin by showing that any point added to the set $\mathcal{U}$ in Algorithm 1 corresponds to a point of $\mathcal{P}$. Consider any iteration of the algorithm where $V^0_{int} + \varepsilon_{int} \cdot A_{int} \leq V^0_L + \varepsilon_{int} \cdot A_L$ and $A_L \neq A_R$. Then, from Lemma 3, it follows that in this case, for any possible subsidy $\varepsilon \in [\varepsilon_L, \varepsilon_R]$, the optimal value function is given by $\bar{U}^{\mathrm{A}}(\pi^\varepsilon; \varepsilon) = \max(V^0_L + \varepsilon \cdot A_L, V^0_R + \varepsilon \cdot A_R)$, with a change in slope at their intersection, *i.e.*, at ε_{int}. That is, ε_{int} is a point of the partition $\mathcal{P}$, and the value $\bar{U}^{\mathrm{S}}(\varepsilon_{int}; \pi_R) = \max_{\varepsilon \in \mathcal{P}(\varepsilon_{int})} \bar{U}^{\mathrm{S}}(\varepsilon; \pi^\varepsilon)$ is stored in the set $\mathcal{U}$.

[21] We denote by $\mathcal{P}(\varepsilon)$ the interval of $\mathcal{P}$ containing ε.

[22] For convenience, we use L, R and int as subscripts instead of using π_L, π_R and π_{int}.

Reciprocally, consider any iteration of the algorithm where $V^0_{int} + \varepsilon_{int} \cdot A_{int} > V^0_L + \varepsilon_{int} \cdot A_L$. Then, Lemma 3 implies that $\mathcal{P}(\varepsilon_L) \neq \mathcal{P}(\varepsilon_{int})$ and $\mathcal{P}(\varepsilon_R) \neq \mathcal{P}(\varepsilon_{int})$. That is, $\mathcal{P}(\varepsilon_{int})$ is a new interval in the partition $\mathcal{P}$ where the optimal value function is given by the linear component $V^0_{\pi_{int}} + \varepsilon \cdot A_{\pi_{int}}$. The algorithm pushes the two sub-intervals $[\varepsilon_L, \varepsilon_{int}]$ and $[\varepsilon_{int}, \varepsilon_R]$ onto the stack (line 1), along with their respective linear components $V^0_L + \varepsilon \cdot A_L$, $V^0_{\pi_{int}} + \varepsilon \cdot A_{\pi_{int}}$ and $V^0_R + \varepsilon \cdot A_R$. Since by Proposition 8, there are a finite number of such linear components (or equivalently, $\mathcal{P}$ is finite), Algorithm 1 terminates in a finite number of steps.

To conclude, we argue that, upon termination, the set $\mathcal{U}$ obtained from Algorithm 1 contains every point of $\mathcal{P}$. Let N denote the total number of pop operations performed by the algorithm before $\mathcal{I}$ becomes empty (finite by the argument above). For $k = 0, 1, \ldots, N$, denote by $\mathcal{I}_k$ the state of the stack after k pops, and by $\mathcal{D}_k \subseteq \mathcal{P}$ the set of points of $\mathcal{P}$ found by the algorithm during the first k iterations, with $\mathcal{D}_0 = \{0, \varepsilon^{\max}\}$. Observe first that whenever Algorithm 1 pushes two new intervals onto $\mathcal{I}$, the intersection point ε_{int} lies strictly interior to the interval $\mathcal{P}(\varepsilon_{int})$ (since by Lemma 3, $\mathcal{P}(\varepsilon_L) \neq \mathcal{P}(\varepsilon_{int}) \neq \mathcal{P}(\varepsilon_R)$, and $\varepsilon_L < \varepsilon_{int} < \varepsilon_R$), so $\varepsilon_{int} \notin \mathcal{P}$. Consequently, every endpoint of an interval ever pushed onto $\mathcal{I}_k$ is either in $\mathcal{D}_k$ or not in $\mathcal{P}$. In light of this, we prove by induction on k the following condition:[23]

$$\mathcal{P} \;\subseteq\; \mathcal{D}_k \;\cup \bigcup_{[\varepsilon_L, \varepsilon_R] \in \mathcal{I}_k} (\varepsilon_L, \varepsilon_R). \tag{60}$$

Base case $k = 0$**.** $\mathcal{I}_0 = \{[0, \varepsilon^{\max}]\}$ and $\mathcal{D}_0 = \{0, \varepsilon^{\max}\}$. Every $\varepsilon \in \mathcal{P}$ is either in $\{0, \varepsilon^{\max}\} = \mathcal{D}_0$ or in $(0, \varepsilon^{\max})$, so the condition in Eq. 60 holds for $k = 0$.

Inductive step $k \to k+1$**.** Suppose the condition in Eq. 60 holds and consider the $(k+1)$-th pop of some $[\varepsilon_L, \varepsilon_R] \in \mathcal{I}_k$.

- If $V^0_{int} + \varepsilon_{int} \cdot A_{int} \leq V^0_L + \varepsilon_{int} \cdot A_L$, then ε_{int} is added to $\mathcal{D}_{k+1}$, and the popped interval is not replaced. By Lemma 3, $(\varepsilon_L, \varepsilon_R) \cap \mathcal{P} \subseteq \{\varepsilon_{int}\} \subseteq \mathcal{D}_{k+1}$. Hence any $\varepsilon \in \mathcal{P}$ previously covered by $(\varepsilon_L, \varepsilon_R)$ is now in $\mathcal{D}_{k+1}$. The endpoints ε_L and ε_R, if they belong to $\mathcal{P}$, were endpoints of the popped interval and hence are already contained in $\mathcal{D}_k \subseteq \mathcal{D}_{k+1}$. Thus, the condition in Eq. 60 holds for $k+1$.
- If $V^0_{int} + \varepsilon_{int} \cdot A_{int} > V^0_L + \varepsilon_{int} \cdot A_L$, then $[\varepsilon_L, \varepsilon_R]$ is replaced by $[\varepsilon_L, \varepsilon_{int}]$ and $[\varepsilon_{int}, \varepsilon_R]$, while $\mathcal{D}_{k+1} = \mathcal{D}_k$. If $\varepsilon_{int} \in \mathcal{D}_k$, Eq. 60 holds for $k+1$, and if $\varepsilon_{int} \notin \mathcal{P}$, we have that $(\varepsilon_L, \varepsilon_R) \cap \mathcal{P} = \big[(\varepsilon_L, \varepsilon_{int}) \cup (\varepsilon_{int}, \varepsilon_R)\big] \cap \mathcal{P}$, and the condition in Eq. 60 also holds for $k+1$.

Lastly, at the final iteration, $\mathcal{I}_N = \emptyset$ implies $\mathcal{P} \subseteq \mathcal{D}_N$; combined with $\mathcal{D}_N \subseteq \mathcal{P}$ by construction, $\mathcal{D}_N = \mathcal{P}$. Since every $\varepsilon \in \mathcal{D}_N$ has its social utility saved in $\mathcal{U}$ (at line 1 for the initial endpoints and line 1 for each new point of $\mathcal{P}$ found), $\mathcal{U}$ contains $\{(\varepsilon, \bar{U}^{\mathrm{S}}(\varepsilon; \pi^{\varepsilon})) : \varepsilon \in \mathcal{P}\}$. The maximizer of $\bar{U}^{\mathrm{S}}$ over $[0, \varepsilon^{\max}]$ is attained at some left endpoint $\varepsilon_i \in \mathcal{P}$, and $\arg\max_{(\varepsilon,u)\in\mathcal{U}} u$ returns it.

[23] With a slight abuse of notation, we identify the elements of $\mathcal{I}$ with intervals, meaning that if $(\varepsilon_L, \pi_L, \bar{V}^0_L, \bar{A}_L, \varepsilon_R, \pi_R, \bar{V}^0_R, \bar{A}_R) \in \mathcal{I}$, we write that $[\varepsilon_L, \varepsilon_R] \in \mathcal{I}$.

G. Additional Experimental Details

The complete code used for our experiments, including the implementation of Algorithm 3 and Algorithm 1, is available at https://github.com/Human-Centric-Machine-Learning/strategic-experiments.

Hardware setup. Our experiments are executed on a compute server equipped with 2 $\times$ Intel Xeon Gold 5317 CPU, 1,024 GB main memory, and 2 $\times$ H100 NVIDIA GPU (80 GB, Hopper Architecture). In each experiment, a single Nvidia H100 GPU is used.

Software setup. All experiments are implemented in `Python` 3.13.5 using `PyTorch` 2.1.1 and `NumPy` 2.4.4. Computations are performed on an NVIDIA GPU with `CUDA` 13.0 support.

Runtime. For the setting used in Section 6, Algorithm 3 computes the optimal policy of the agent (for any given subsidy) in $\sim$ 2.4s, and Algorithm 1 computes the optimal subsidy in $\sim$ 271s, which involves solving 114 times a different belief MDP $\mathcal{M}^{\varepsilon}$. In Figure 4, we evaluate the runtime of Algorithm 1 across multiple configurations for the parameters T and $n^{\max}$, which determine the size of the state and action space (Proposition 4).

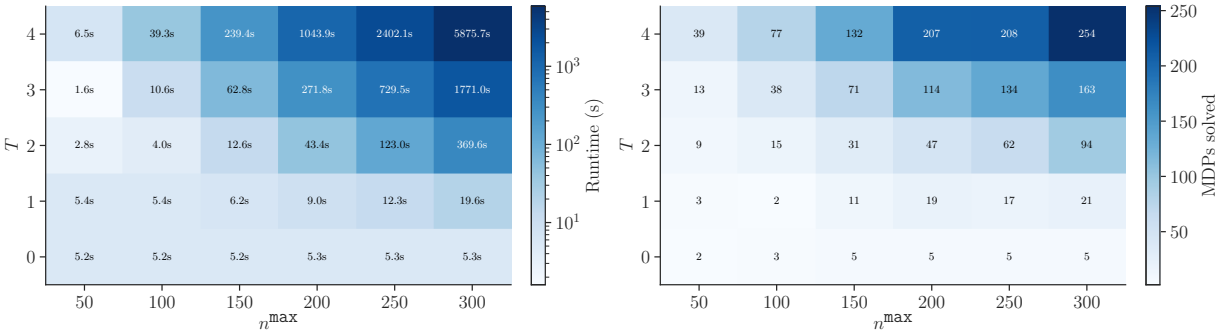


Figure 4. **Runtime of Algorithm 1.** The figure shows, across multiple values of the maximum number of trials T and the maximum sample size per trial $n^{\max}$, the runtime of Algorithm 1 (left panel) and the number of belief MDPs solved by the algorithm (right panel). All other parameters are fixed as specified in Tables 2 and 3. The experiments are run on an NVIDIA H100 GPU.

Parameter details. Tables 2 and 3 report the values of all parameters required to specify the sequential approval protocol used in our fiducial setting (Section 6). Unless otherwise stated, all results are obtained using these values. When any parameter is varied (*e.g.*, in Panel (b) of Figure 1 or in Appendix H), we explicitly indicate it.

Table 2. **Non-economic parameters**

T	$n^{\max}$	θ^{b}	κ	θ^{*}	$\varepsilon^{\max}$	(α_0, β_0)	Q
3	200	0.5	0.05	0.65	0.9	(1,1)	$\delta_{(1,1)}$

Table 3. **Economic parameters**

ρ^{S}	ρ^{A}	$c(n) = c_0 + c_1 n,\ n \neq 0$	
		c_0	c_1
\$2000 M	\$240 M	\$48.9 M	\$0.066 M

Implementation details. Our Python implementation of Algorithm 3 leverages the bijection between pairs (α, β) and pairs (X, N), where X is the total number of successes and N the total number of patients (the total sample size), as we detail in Eq. 33. Furthermore, since the cost function is linear, we also use the fact that the total cumulated cost C at a given state (X, N) at time l can be written as $C = l \cdot c_0 + N \cdot c_1$. As a result, our implementation does not explicitly keep track of the cumulated cost C, which significantly reduces the computational overhead. A similar idea is used in the proof of Proposition 4.

Our implementation of Algorithm 1 computes the optimal subsidy ε^* (and solves each belief MDP $\mathcal{M}^\varepsilon$) exactly. Consequently, there is no associated uncertainty in ε^* or in any of the quantities that we report computed using the belief MDP, including $\bar{U}^{\mathbb{A}}$ and $\bar{U}^{\mathbb{S}}$. In contrast, the true quantities under the approval process using θ^* described in Section 2—such as the true utilities $U^{\mathbb{A}}$ and $U^{\mathbb{S}}$ (Eq. 7), as well as the probability that the agent opts out during the approval process—are estimated using $100{,}000 - 200{,}000$ Monte Carlo rollouts, and we report 95% confidence error bars computed via bootstrapping with 1000 resamples.

H. Additional Experimental Results

H.1. Additional results complementing Section 6

In this section, we provide complementary results to the approval process considered in Section 6, whose parameters are given in Table 2 and Table 3.

In Figure 5, we show the optimal value function and policy for the MDP $\mathcal{M}^{\varepsilon^*}$, with $\varepsilon^* = 1.08$, and in Figure 6 we show how the belief of the agent evolves in 300 realizations of the approval process, for different values of the true efficacy θ^*.

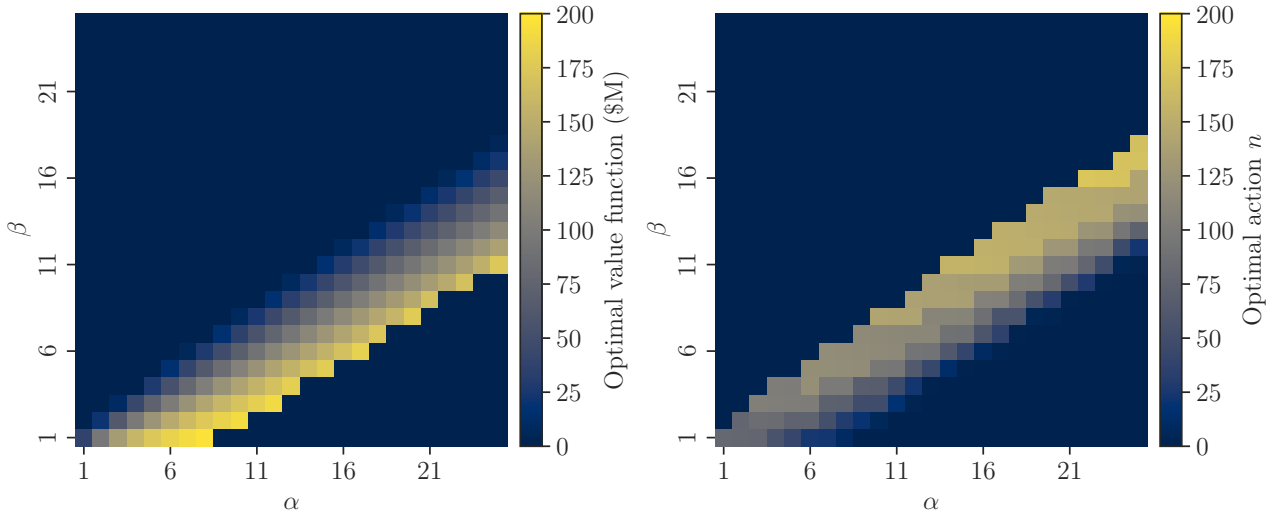
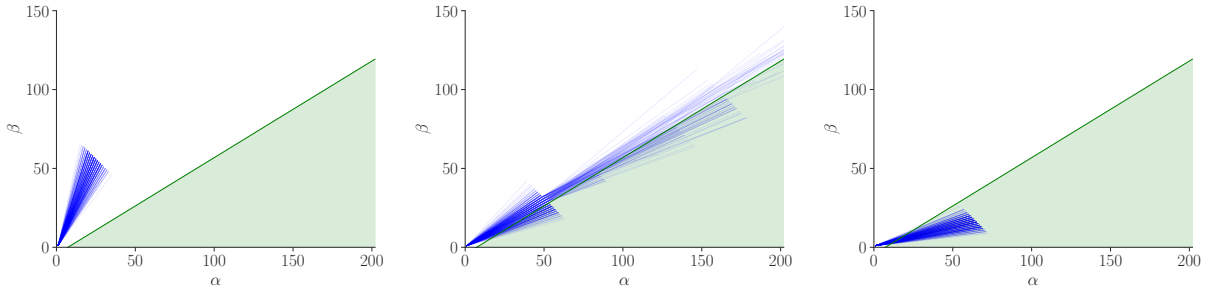


Figure 5. **Optimal value function and policy in the belief MDP $\mathcal{M}^{\varepsilon^*}$ for the optimal subsidy $\varepsilon^* = 0.108$.** The left panel shows the optimal value function in the belief MDP, $V^{\varepsilon^*}(\alpha, \beta, C(\alpha, \beta), 1)$, at time step $l = 1$, where the cost of each state is given by $C(\alpha, \beta) = 1 \cdot c_0 + (\alpha + \beta - \alpha_0 - \beta_0) \cdot c_1$ (see Eq. 33). The right panel shows the optimal action n taken by the optimal policy at time step $l = 1$ for each belief, *i.e.*, $\pi^{\varepsilon^*}(\alpha, \beta, C(\alpha, \beta), 1)$. The optimal action at time step $l = 0$ (not shown here) is $n = 79$.



(a) $\theta^* = 0.3$ *(b)* $\theta^* = 0.65$ *(c)* $\theta^* = 0.8$

Figure 6. **Trajectories of the approval process.** Each panel shows 300 realizations of the approval process for different true efficacies θ^* of the antibiotic. Each blue segment corresponds to the agent conducting a new trial and updating its belief (see Figure 3 for an illustration of the geometry). In the left panel, the agent conducts a first trial and then always opts out; in the middle panel, the agent can conduct multiple trials, and in the right panel, the antibiotic is always approved after the first trial.

In Figure 7, we show, for the initial action taken by the agent at time step $l = 0$ (and belief $(\alpha_0, \beta_0) = (1, 1)$), the expected cumulative future reward in $\mathcal{M}^{\varepsilon^*}$ for the optimal subsidy $\varepsilon^* = 0.108$, defined for each $n > 0$ as $\mathbb{E}_{\pi^{\varepsilon^*}}[\sum_{t=0}^{T} r^{\varepsilon^*}(S_t, n_t, S_{t+1}) | S_0, n_0 = n]$. The sample size maximizing this curve is $n = 79$, which is the size of the first trial conducted by the agent. Notably, although the expected reward is unimodal, it exhibits small-scale oscillations. These oscillations are not numerical artifacts. In fact, their approximate period is given by $1/\log(1 + \theta^{\flat}(e - 1))$. The reason is that the agent's actions are discrete, whereas the function f defined in Proposition 2, which determines the approval condition, decreases by exactly $\log(1 + \theta^{\flat}(e - 1))$ whenever the agent selects a new action n (recall that

$\alpha_{t+1} + \beta_{t+1} - \alpha_t - \beta_t = n_t$). Since this change is not an integer quantity, the expected reward only exhibits a small decrease after n increases by approximately $1/\log(1 + \theta^{\flat}(e-1))$, which gives rise to the observed oscillations.

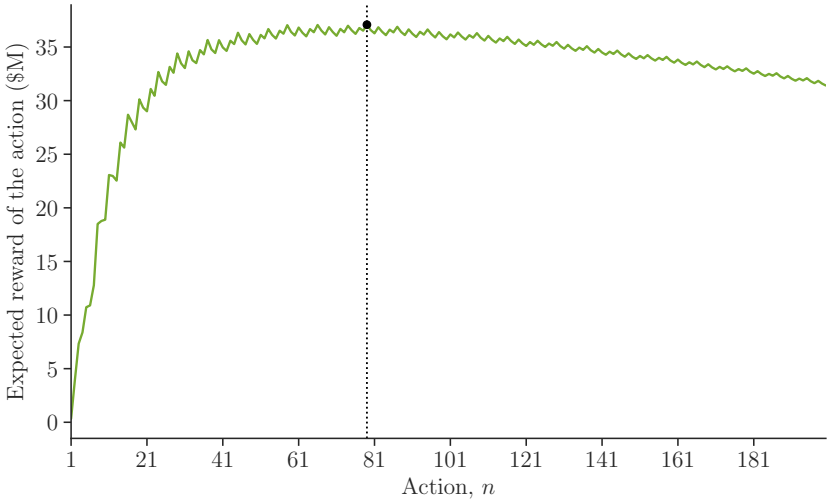
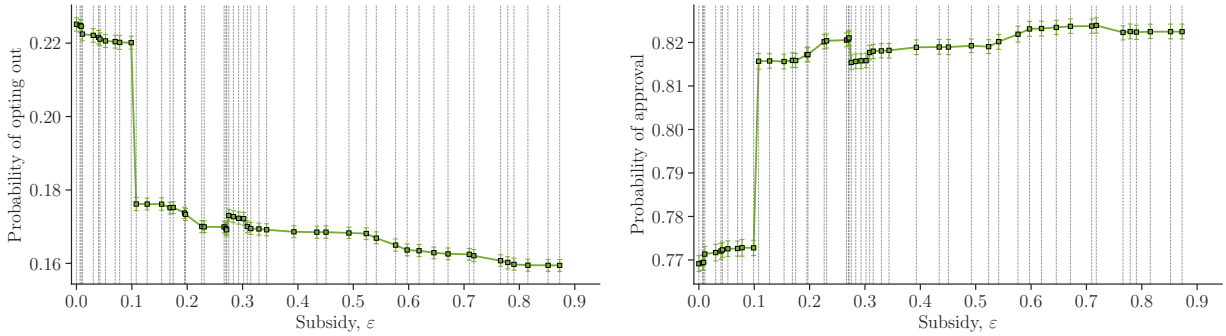


Figure 7. **Expected reward for each sample size.** The figure show, for the initial action taken by the agent at time step $l = 0$ and state $(\alpha_0 = 1, \beta_0 = 1, 0)$, the total expected reward in the MDP $\mathcal{M}^{\varepsilon^*}$ under the optimal subsidy $\varepsilon^* = 0.108$ when the agent takes action n and then follows the optimal policy π^{ε^*}.



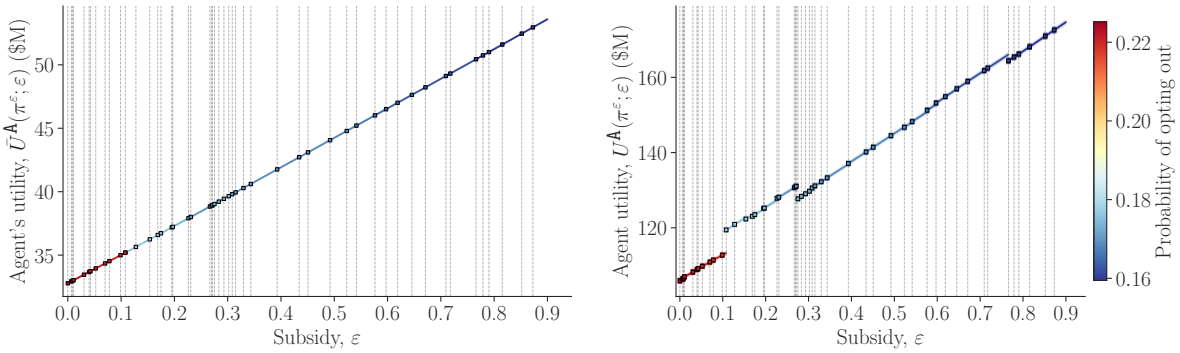
(a) Opting out before approval vs. subsidy

(b) Approval probability vs subsidy

Figure 8. **Opt out and approval probabilities.** The figure shows, for an antibiotic with $\theta^* = 0.65$, the probability that the agent opts out of the approval process by selecting $n = 0$ before approval, as well as the probability that the antibiotic is ultimately approved. For each subsidy level, the agent follows the optimal policy. Note that, in principle, the agent may never opt out during the approval process; however, the antibiotic may still fail to be approved if the maximum number of trials is reached.

In Figure 9 and Figure 10, we show, respectively, the utility of the agent and the social utility when the agent selects its optimal policy for each possible subsidy, both computed using the belief MDP, and the true realized utilities when the antibiotic has efficacy $\theta^* = 0.65$ (Eq. 7).

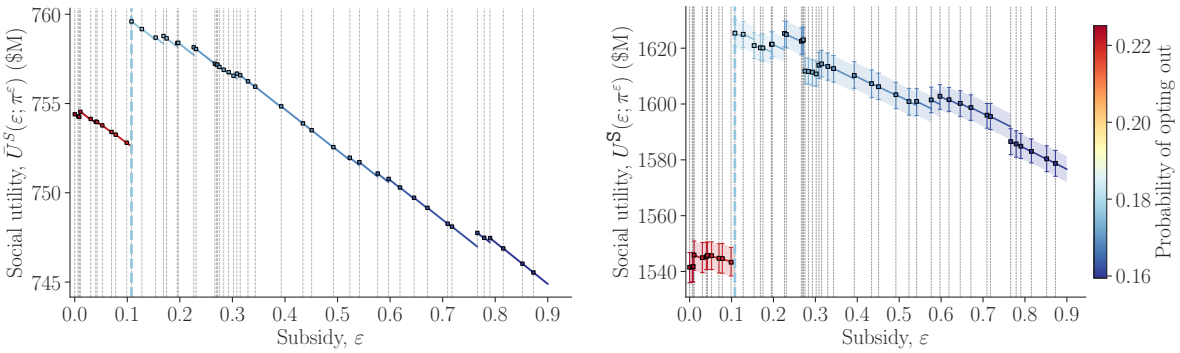
Lastly, in Figure 11 we show that the optimal social utility (that is, the social utility under the optimal subsidy) increases monotonically as the true efficacy θ^* increases.



(a) Agent utility $\bar{U}^{\mathrm{A}}(\pi^{\varepsilon};\varepsilon)$ computed using $\mathcal{M}^{\varepsilon}$

(b) Agent utility $U^{\mathrm{A}}(\pi^{\varepsilon};\varepsilon)$ in the approval process

Figure 9. **Agent utilities.** The left panel shows the agent's utility (Eq. 10) computed using the belief MDP $\mathcal{M}^{\varepsilon}$ when the agent uses the optimal policy for each subsidy, which is a piece-wise linear, convex and continuous function in accordance with Proposition 8. The right panel shows the true utility of the agent (Eq. 7) in the approval process when using the optimal policy π^{ε} for each subsidy and $\theta^* = 0.65$. The dashed vertical lines correspond to the intervals of the partition $\mathcal{P}$ where the agent's optimal policy is constant (Proposition 8).



(a) Social utility $\bar{U}^{\mathrm{S}}(\varepsilon;\pi^{\varepsilon})$ computed using $\mathcal{M}^{\varepsilon}$

(b) Social utility $U^{\mathrm{S}}(\varepsilon;\pi^{\varepsilon})$ in the approval process

Figure 10. **Social utilities.** The left panel shows the social utility (Eq. 13) computed using the belief MDP $\mathcal{M}^{\varepsilon}$ when the agent uses the optimal policy for each subsidy. The right panel shows the true social utility (Eq. 7) in the approval process when the agent uses the optimal policy π^{ε} for each subsidy and $\theta^* = 0.65$. The dashed vertical lines correspond to the intervals of the partition $\mathcal{P}$ where the agent's optimal policy is constant (Proposition 8).

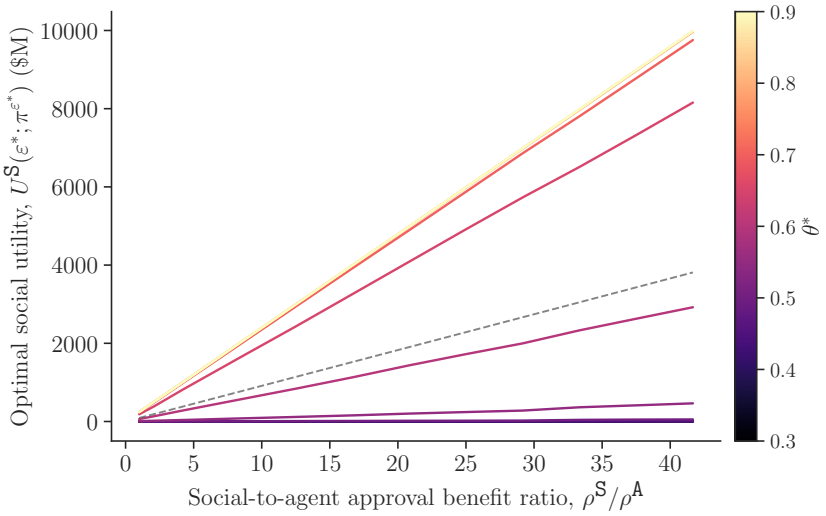


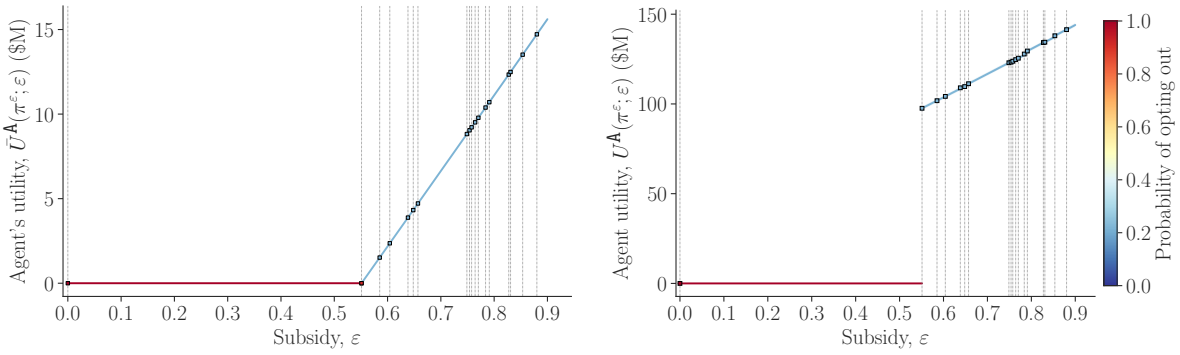
Figure 11. **Optimal social utility for different antibiotic efficacies.** The figure shows how social utility $U^{\mathtt{S}}(\varepsilon^*; \pi^{\varepsilon^*})$—when the principal chooses the optimal subsidy ε^* and the agent adopts the corresponding optimal policy π^{ε^*}—varies as a function of the ratio $\rho^{\mathtt{S}}/\rho^{\mathtt{A}}$ across different levels of efficacy θ^*. The dashed line corresponds to the social utility $\bar{U}^{\mathtt{S}}(\varepsilon^*; \pi^{\varepsilon^*})$ computed using the belief MDP, which does not depend on the true efficacy θ^*.

H.2. Additional results using different parameters

In this section, we present further experimental results for the antibiotic approval process described in Section 6, where we vary selected parameters (see Tables 2 and 3).

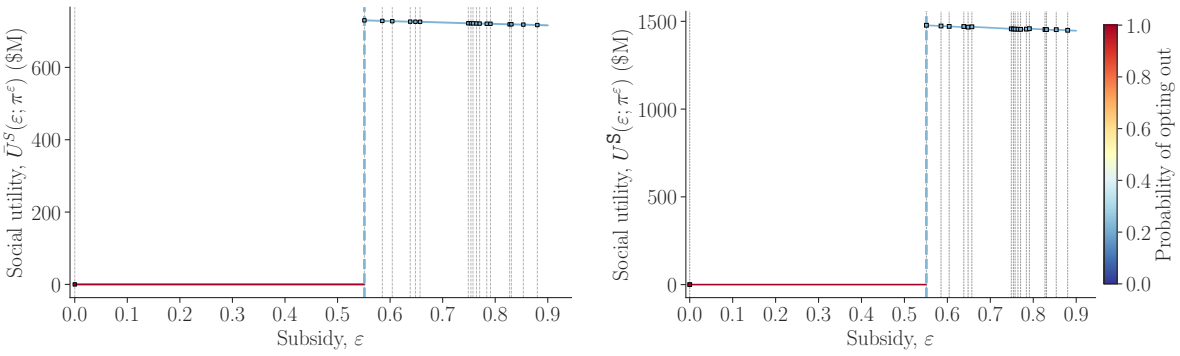
H.2.1. Approval under increased experimental costs

Here, we show the result of an antibiotic approval process with an increased experimental cost. In particular, we take the parameters in Table 2 and Table 3 but increase the fixed cost c_0 of a trial to \$100 M and the per-patient cost c_1 to \$0.1 M. We find that the agent opts out at the beginning of the process unless the principal subsidizes a fraction higher than the optimal subsidy, $\varepsilon^* = 0.551$.



(a) Agent utility $\bar{U}^{\mathrm{A}}(\pi^\varepsilon;\varepsilon)$ computed using $\mathcal{M}^\varepsilon$ *(b)* Agent utility $U^{\mathrm{A}}(\pi^\varepsilon;\varepsilon)$ in the approval process

Figure 12. **Agent utilities under increased experimental costs.** The left panel shows the agent's utility (Eq. 10) computed using the belief MDP $\mathcal{M}^\varepsilon$ when the agent uses the optimal policy for each subsidy, which is a piece-wise linear, convex, and continuous function in accordance with Proposition 8. The right panel shows the true utility of the agent (Eq. 7) in the approval process when using the optimal policy π^ε for each subsidy and $\theta^* = 0.65$. The dashed vertical lines correspond to the intervals of the partition $\mathcal{P}$ where the agent's optimal policy is constant (Proposition 8).



(a) Social utility $\bar{U}^{\mathrm{S}}(\varepsilon;\pi^\varepsilon)$ computed using $\mathcal{M}^\varepsilon$ *(b)* Social utility $U^{\mathrm{S}}(\varepsilon;\pi^\varepsilon)$ in the approval process

Figure 13. **Social utilities under increased experimental costs.** The left panel shows the social utility (Eq. 13) computed using the belief MDP $\mathcal{M}^\varepsilon$ when the agent uses the optimal policy for each subsidy. The right panel shows the true social utility (Eq. 7) in the approval process when the agent uses the optimal policy π^ε for each subsidy and $\theta^* = 0.65$. The dashed vertical lines correspond to the intervals of the partition $\mathcal{P}$ where the agent's optimal policy is constant (Proposition 8).

In Figure 14 we show that this optimal subsidy is constant as ρ^{S} increases, and in Figure 15 that our sequential protocol yields social utility gains $> 20\%$ relative to a non-sequential protocol.

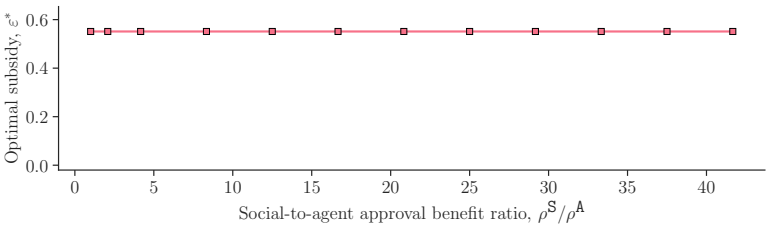


Figure 14. **Optimal subsidy vs. $\rho^{\mathtt{S}}/\rho^{\mathtt{A}}$.** The figure shows, as a function of the social-to-agent approval benefit ratio, the optimal subsidy obtained using Algorithm 1.

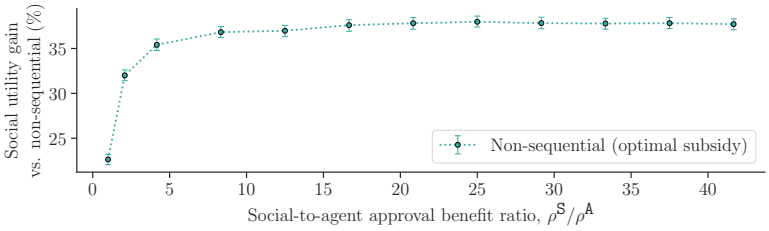
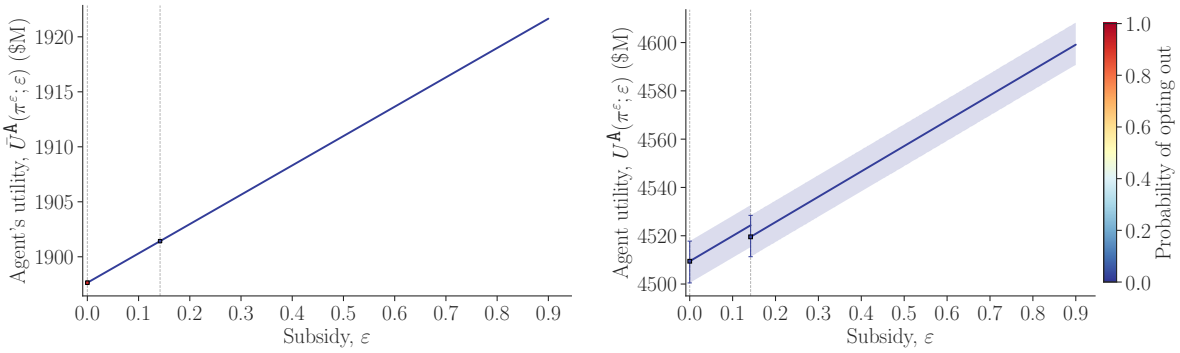


Figure 15. **Social utility gain vs. $\rho^{\mathtt{S}}/\rho^{\mathtt{A}}$.** The figure shows, as a function of the social-to-agent approval benefit ratio, the percentage increase in social utility of the sequential approval protocol relative to a non-sequential approval protocol in which the agent is restricted to a single trial with $n^{\mathtt{max}} = 800$, under the optimal subsidy computed using Algorithm 1 (in the non-sequential protocol without subsidy the agent always opts out, yielding zero social utility).

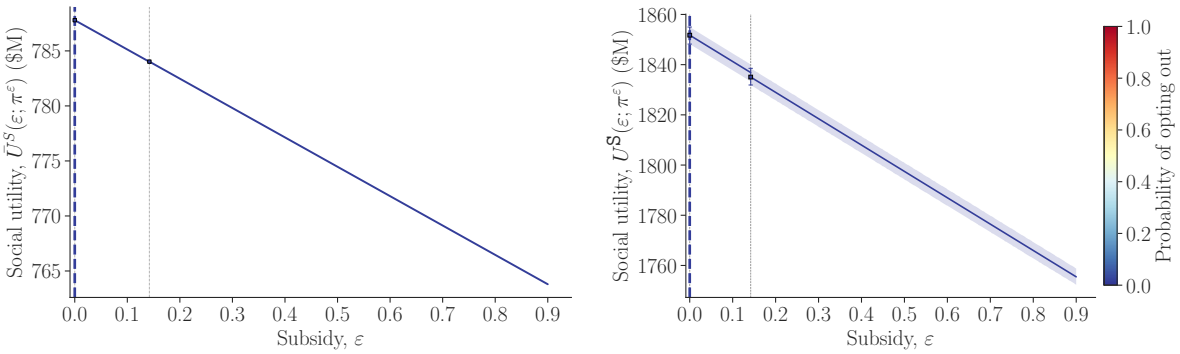
H.2.2. Approval under increased agent approval benefit

Here, we show the result of an antibiotic approval process with an increased approval utility for the agent, $\rho^{\mathbb{A}} = \$5000\,\text{M}$ (and the rest of the parameters in Table 2 and Table 3 fixed). In this case, the approval utility for the agent covers the expected cost of approval by a large margin, and we find that no subsidy is needed, namely, $\varepsilon^* = 0$.



(a) Agent utility $\bar{U}^{\mathbb{A}}(\pi^\varepsilon;\varepsilon)$ computed using $\mathcal{M}^\varepsilon$ *(b)* Agent utility $U^{\mathbb{A}}(\pi^\varepsilon;\varepsilon)$ in the approval process

Figure 16. **Agent utilities under increased approval utility.** The left panel shows the agent's utility (Eq. 10) computed using the belief MDP $\mathcal{M}^\varepsilon$ when the agent uses the optimal policy for each subsidy, which is a piece-wise linear, convex and continuous function in accordance with Proposition 8. The right panel shows the true utility of the agent (Eq. 7) in the approval process when using the optimal policy π^ε for each subsidy and $\theta^* = 0.65$. The dashed vertical lines correspond to the intervals of the partition $\mathcal{P}$ where the agent's optimal policy is constant (Proposition 8).



(a) Social utility $\bar{U}^{\mathbb{S}}(\varepsilon;\pi^\varepsilon)$ computed using $\mathcal{M}^\varepsilon$ *(b)* Social utility $U^{\mathbb{S}}(\varepsilon;\pi^\varepsilon)$ in the approval process

Figure 17. **Social utilities under increased approval agent utility.** The left panel shows the social utility (Eq. 13) computed using the belief MDP $\mathcal{M}^\varepsilon$ when the agent uses the optimal policy for each subsidy. The right panel shows the true social utility (Eq. 7) in the approval process when the agent uses the optimal policy π^ε for each subsidy and $\theta^* = 0.65$. The dashed vertical lines correspond to the intervals of the partition $\mathcal{P}$ where the agent's optimal policy is constant (Proposition 8).

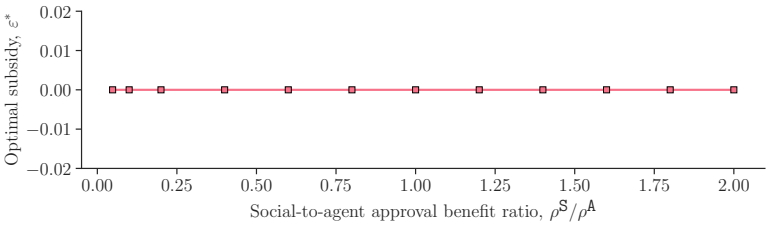


Figure 18. **Optimal subsidy vs. $\rho^{\mathbb{S}}/\rho^{\mathbb{A}}$.** The figure shows, as a function of the social-to-agent approval benefit ratio, the optimal subsidy obtained using Algorithm 1.

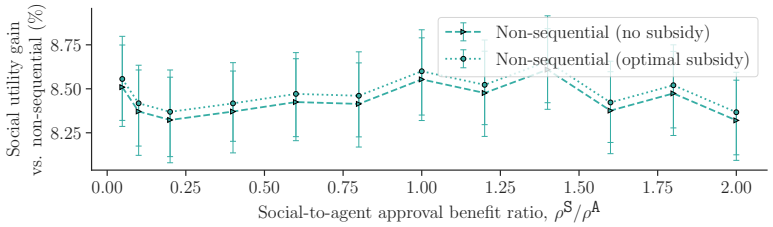
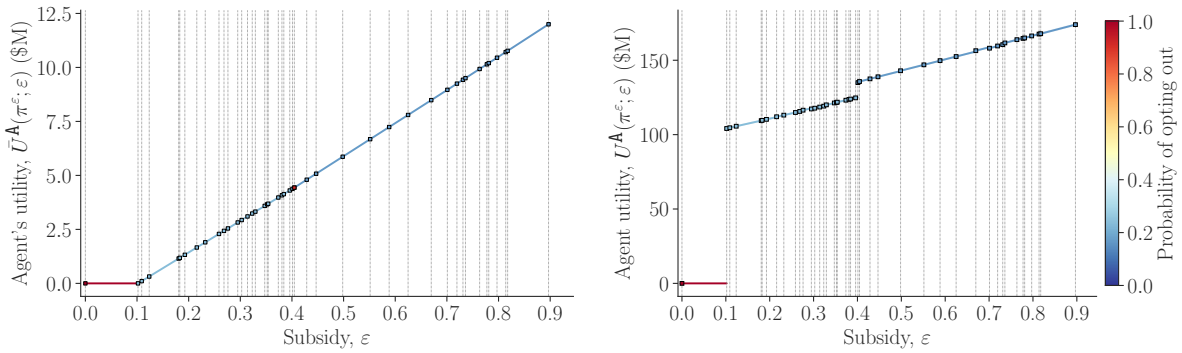


Figure 19. **Social utility gain vs. $\rho^{\mathtt{S}}/\rho^{\mathtt{A}}$.** The figure shows, as a function of the social-to-agent approval benefit ratio, the percentage increase in social utility of the sequential approval protocol relative to a non-sequential approval protocol in which the agent is restricted to a single trial with $n^{\max} = 800$, under (i) the optimal subsidy computed using Algorithm 1 and (ii) no subsidy ($\varepsilon = 0$). In this case, the optimal non-sequential policy is the same for all subsidies, and we apply a small vertical jitter to improve visibility.

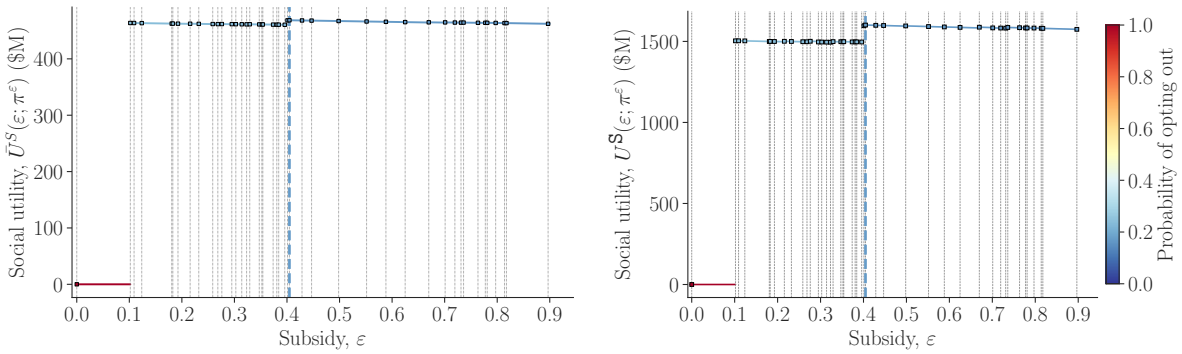
H.2.3. APPROVAL UNDER PESSIMISTIC PRIOR

Here, we show the results of an antibiotic approval process where the agent's prior is $(\alpha_0, \beta_0) = (1, 1.5)$, that is, the agent is slightly pessimistic about its product.[24] Note that the mean efficacy of the drug according to its prior is then $0.4 < \theta^* = 0.65$. The principal knows such prior, and the rest of the parameters in Table 2 and Table 3 are fixed. In this case, we find that the agent opts out at the beginning of the approval process as long as the subsidy is $\varepsilon \lesssim 0.1$. The optimal subsidy is $\varepsilon^* \approx 0.4$.



(a) Agent utility $\bar{U}^{\mathbb{A}}(\pi^\varepsilon;\varepsilon)$ computed using $\mathcal{M}^\varepsilon$ *(b)* Agent utility $U^{\mathbb{A}}(\pi^\varepsilon;\varepsilon)$ in the approval process

Figure 20. **Agent utilities under a pessimistic prior.** The left panel shows the agent's utility (Eq. 10) computed using the belief MDP $\mathcal{M}^\varepsilon$ when the agent uses the optimal policy for each subsidy, which is a piece-wise linear, convex and continuous function in accordance with Proposition 8. The right panel shows the true utility of the agent (Eq. 7) in the approval process when using the optimal policy π^ε for each subsidy and $\theta^* = 0.65$. The dashed vertical lines correspond to the intervals of the partition $\mathcal{P}$ where the agent's optimal policy is constant (Proposition 8).



(a) Social utility $\bar{U}^{\mathbb{S}}(\varepsilon;\pi^\varepsilon)$ computed using $\mathcal{M}^\varepsilon$ *(b)* Social utility $U^{\mathbb{S}}(\varepsilon;\pi^\varepsilon)$ in the approval process

Figure 21. **Social utilities under a pessimistic prior.** The left panel shows the social utility (Eq. 13) computed using the belief MDP $\mathcal{M}^\varepsilon$ when the agent uses the optimal policy for each subsidy. The right panel shows the true social utility (Eq. 7) in the approval process when the agent uses the optimal policy π^ε for each subsidy and $\theta^* = 0.65$. The dashed vertical lines correspond to the intervals of the partition $\mathcal{P}$ where the agent's optimal policy is constant (Proposition 8).

In Figure 22 we show that the optimal subsidy increases with $\rho^{\mathbb{S}}/\rho^{\mathbb{A}}$, and in Figure 23 that our sequential protocol yields gains $> 40\%$ in social utility relative to a non-sequential protocol.

[24] We select $\beta_0 = 1.5$ as higher values lead the agent to opt out at the beginning of the process, regardless of the subsidy.

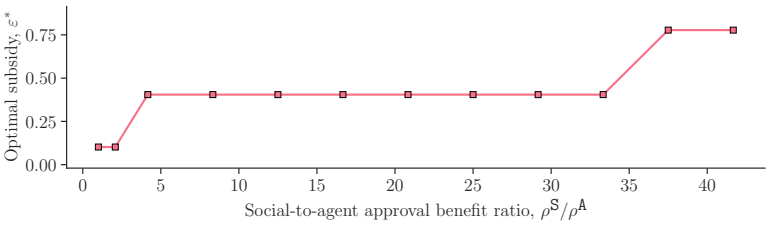


Figure 22. **Optimal subsidy vs. $\rho^{\mathtt{S}}/\rho^{\mathtt{A}}$.** The figure shows, as a function of the social-to-agent approval benefit ratio, the optimal subsidy obtained using Algorithm 1.

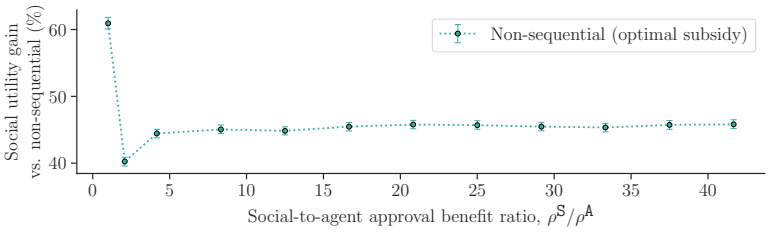
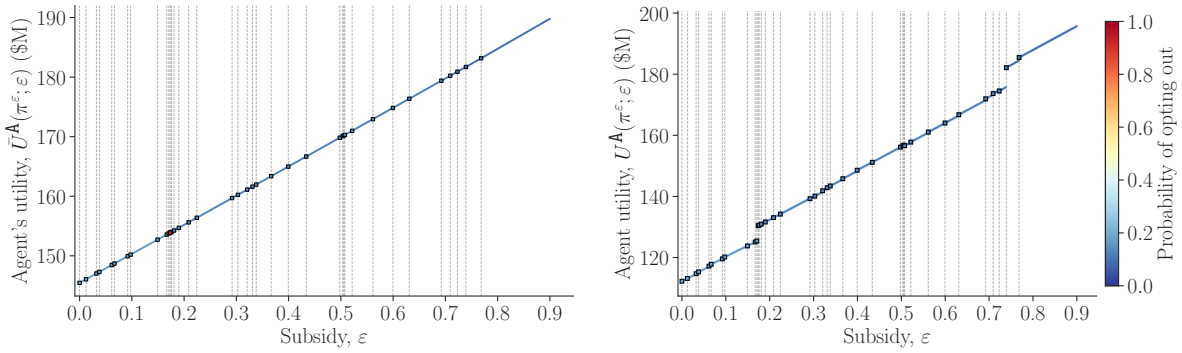


Figure 23. **Social utility gain vs. $\rho^{\mathtt{S}}/\rho^{\mathtt{A}}$.** The figure shows, as a function of the social-to-agent approval benefit ratio, the percentage increase in social utility of the sequential approval protocol relative to a non-sequential approval protocol in which the agent is restricted to a single trial with $n^{\max} = 800$, under (i) the optimal subsidy computed using Algorithm 1 (in the non-sequential protocol without subsidy the agent always opts out, yielding zero social utility).

H.2.4. APPROVAL UNDER OPTIMISTIC PRIOR

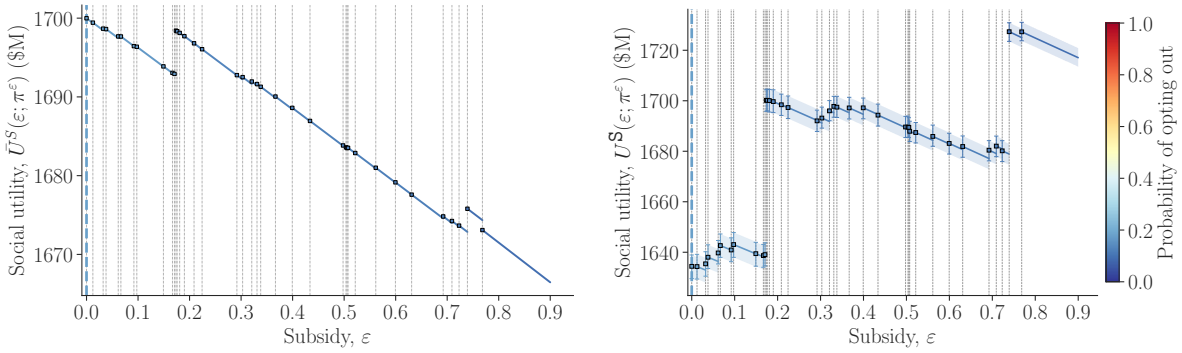
Here, we show the results of an antibiotic approval process where the agent's prior is $(\alpha_0, \beta_0) = (4, 1)$, that is, the agent is slightly optimistic about the antibiotic. Note that the mean efficacy of the drug according to its prior is then $0.8 > \theta^* = 0.65$. The principal knows such prior, and the rest of the parameters in Table 2 and Table 3 are fixed. In this case, the optimal subsidy is $\varepsilon^* = 0$, and as can be seen in Figure 25, the optimal subsidy maximizing $\bar{U}^{\mathrm{S}}(\varepsilon; \pi^\varepsilon)$ does not necessarily maximize the true (unknown) utility $U^{\mathrm{S}}(\varepsilon; \pi^\varepsilon)$. Nevertheless, Figure 27 shows that the proposed sequential subsidized protocol can still substantially improve social utility relative to a non-sequential protocol, with gains exceeding approximately 18%.



(a) Agent utility $\bar{U}^{\mathrm{A}}(\pi^\varepsilon; \varepsilon)$ computed using $\mathcal{M}^\varepsilon$

(b) Agent utility $U^{\mathrm{A}}(\pi^\varepsilon; \varepsilon)$ in the approval process

Figure 24. **Agent utilities under an optimist prior.** The left panel shows the agent's utility (Eq. 10) computed using the belief MDP $\mathcal{M}^\varepsilon$ when the agent uses the optimal policy for each subsidy, which is a piece-wise linear, convex and continuous function in accordance with Proposition 8. The right panel shows the true utility of the agent (Eq. 7) in the approval process when using the optimal policy π^ε for each subsidy and $\theta^* = 0.65$. The dashed vertical lines correspond to the intervals of the partition $\mathcal{P}$ where the agent's optimal policy is constant (Proposition 8).



(a) Social utility $\bar{U}^{\mathrm{S}}(\varepsilon; \pi^\varepsilon)$ computed using $\mathcal{M}^\varepsilon$

(b) Social utility $U^{\mathrm{S}}(\varepsilon; \pi^\varepsilon)$ in the approval process

Figure 25. **Social utilities under an optimist prior.** The left panel shows the social utility (Eq. 13) computed using the belief MDP $\mathcal{M}^\varepsilon$ when the agent uses the optimal policy for each subsidy. The right panel shows the true social utility (Eq. 7) in the approval process when the agent uses the optimal policy π^ε for each subsidy and $\theta^* = 0.65$. The dashed vertical lines correspond to the intervals of the partition $\mathcal{P}$ where the agent's optimal policy is constant (Proposition 8).

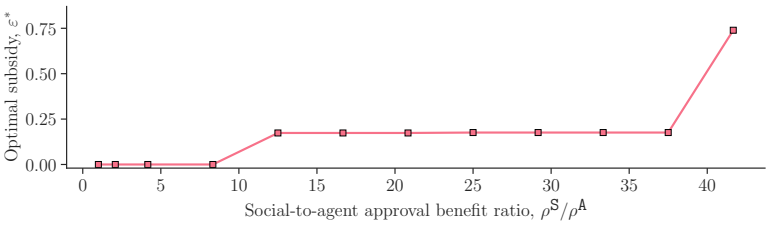


Figure 26. **Optimal subsidy vs. $\rho^{\mathtt{S}}/\rho^{\mathtt{A}}$.** The figure shows, as a function of the social-to-agent approval benefit ratio, the optimal subsidy obtained using Algorithm 1.

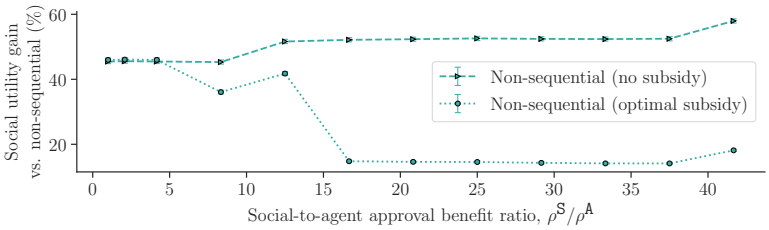
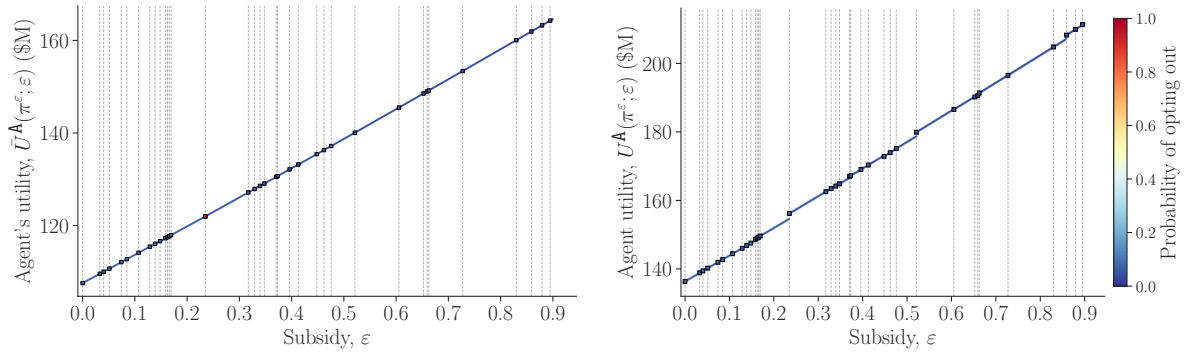


Figure 27. **Social utility gain vs. $\rho^{\mathtt{S}}/\rho^{\mathtt{A}}$.** The figure shows, as a function of the social-to-agent approval benefit ratio, the percentage increase in social utility of the sequential approval protocol relative to a non-sequential approval protocol in which the agent is restricted to a single trial with $n^{\max} = 800$, under (i) the optimal subsidy computed using Algorithm 1 and (ii) no subsidy ($\varepsilon = 0$).

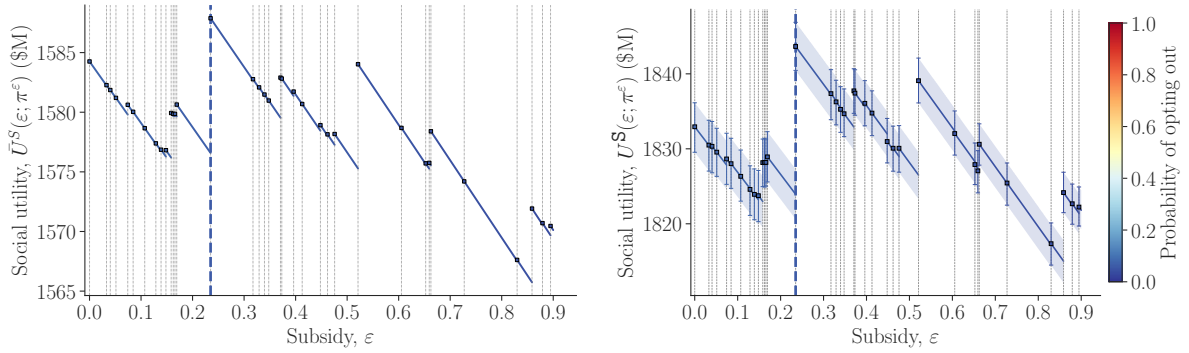
H.2.5. APPROVAL UNDER CALIBRATED PRIOR

Here, we show the results of an antibiotic approval process where the agent's prior is $(\alpha_0, \beta_0) = (130, 70)$. This corresponds to a very informative prior that is calibrated to the true efficacy $\theta^* = 0.65$, since the mean of the prior is precisely 0.65. The principal knows such prior, and the rest of the parameters in Table 2 and Table 3 are fixed. In this case, we find that the calibrated prior allows the agent to increase its utility $U^{\mathbb{A}}(\pi^\varepsilon; \varepsilon)$ for any possible subsidy, as can be seen by comparing Figure 28 to the fiducial setting in Figure 9. However, and perhaps surprisingly, we also find that it is still optimal for the principal to subsidize a non-negligible fraction $\varepsilon^* \approx 0.234$ of the agent's cost, despite the prior belief supporting that the drug should be approved.



(a) Agent utility $\bar{U}^{\mathbb{A}}(\pi^\varepsilon; \varepsilon)$ computed using $\mathcal{M}^\varepsilon$ *(b)* Agent utility $U^{\mathbb{A}}(\pi^\varepsilon; \varepsilon)$ in the approval process

Figure 28. **Agent utilities under a calibrated informative prior.** The left panel shows the agent's utility (Eq. 10) computed using the belief MDP $\mathcal{M}^\varepsilon$ when the agent uses the optimal policy for each subsidy, which is a piece-wise linear, convex and continuous function in accordance with Proposition 8. The right panel shows the true utility of the agent (Eq. 7) in the approval process when using the optimal policy π^ε for each subsidy and $\theta^* = 0.65$. The dashed vertical lines correspond to the intervals of the partition $\mathcal{P}$ where the agent's optimal policy is constant (Proposition 8).



(a) Social utility $\bar{U}^{\mathrm{S}}(\varepsilon; \pi^\varepsilon)$ computed using $\mathcal{M}^\varepsilon$ *(b)* Social utility $U^{\mathrm{S}}(\varepsilon; \pi^\varepsilon)$ in the approval process

Figure 29. **Social utilities under a calibrated informative prior.** The left panel shows the social utility (Eq. 13) computed using the belief MDP $\mathcal{M}^\varepsilon$ when the agent uses the optimal policy for each subsidy. The right panel shows the true social utility (Eq. 7) in the approval process when the agent uses the optimal policy π^ε for each subsidy and $\theta^* = 0.65$. The dashed vertical lines correspond to the intervals of the partition $\mathcal{P}$ where the agent's optimal policy is constant (Proposition 8).

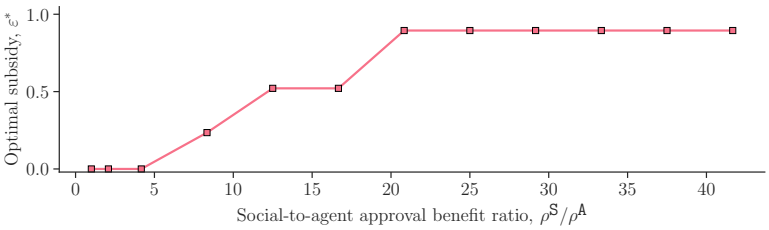


Figure 30. **Optimal subsidy vs. $\rho^{\mathtt{S}}/\rho^{\mathtt{A}}$.** The figure shows, as a function of the social-to-agent approval benefit ratio, the optimal subsidy obtained using Algorithm 1.

Lastly, Figure 31 shows that the social utility gain relative to a non-sequential protocol with optimal subsidies is small for high values of $\rho^{\mathtt{S}}/\rho^{\mathtt{A}}$. This is expected because under a calibrated prior, the agent already has an accurate estimate of the product's efficacy and therefore gathering new information brings little benefit. As a result, the agent can effectively select an optimal sample size and complete the process in a single trial.

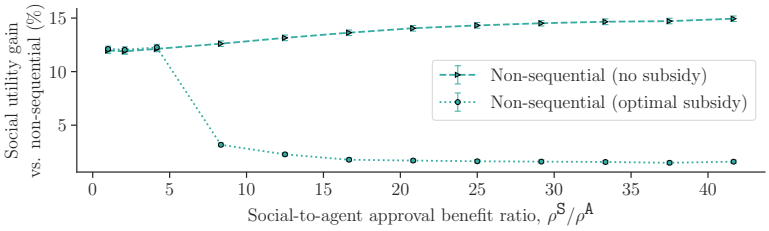
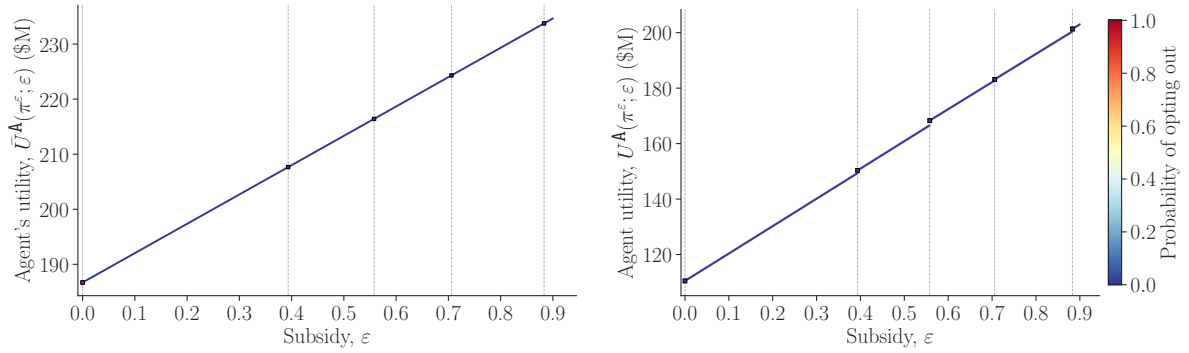


Figure 31. **Social utility gain vs. $\rho^{\mathtt{S}}/\rho^{\mathtt{A}}$.** The figure shows, as a function of the social-to-agent approval benefit ratio, the percentage increase in social utility of the sequential approval protocol relative to a non-sequential approval protocol in which the agent is restricted to a single trial with $n^{\max} = 800$, under (i) the optimal subsidy computed using Algorithm 1 and (ii) no subsidy ($\varepsilon = 0$).

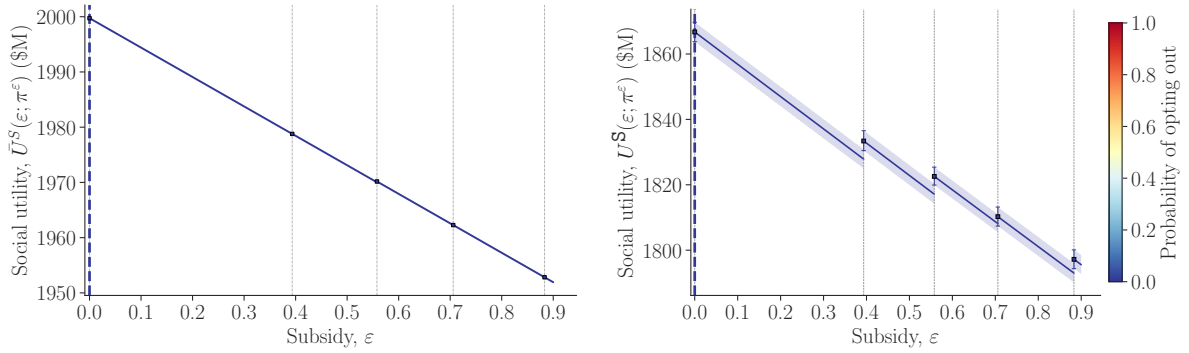
H.2.6. Approval under uncalibrated prior

Here, we show the results of an antibiotic approval process where the agent's prior is $(\alpha_0, \beta_0) = (130, 30)$. This corresponds to a very informative and optimistic prior that is uncalibrated to the true efficacy $\theta^* = 0.65$, since the mean of the prior is precisely 0.8125. The principal knows such prior, and the rest of the parameters in Table 2 and Table 3 are fixed. We find that, across subsidies, the utility $U^{\mathbb{A}}(\pi^\varepsilon; \varepsilon)$ achieved by the agent slightly decreased compared to an agent with a calibrated prior (Figure 32 vs. Figure 28). However, surprisingly, we also find that the social utility is greater in this case compared to the case where the agent has a calibrated prior (Figure 33 vs. Figure 29).



(a) Agent utility $\bar{U}^{\mathbb{A}}(\pi^\varepsilon; \varepsilon)$ computed using $\mathcal{M}^\varepsilon$ *(b)* Agent utility $U^{\mathbb{A}}(\pi^\varepsilon; \varepsilon)$ in the approval process

Figure 32. **Agent utilities under an uncalibrated informative prior.** The left panel shows the agent's utility (Eq. 10) computed using the belief MDP $\mathcal{M}^\varepsilon$ when the agent uses the optimal policy for each subsidy, which is a piece-wise linear, convex and continuous function in accordance with Proposition 8. The right panel shows the true utility of the agent (Eq. 7) in the approval process when using the optimal policy π^ε for each subsidy and $\theta^* = 0.65$. The dashed vertical lines correspond to the intervals of the partition $\mathcal{P}$ where the agent's optimal policy is constant (Proposition 8).



(a) Social utility $\bar{U}^{\mathrm{S}}(\varepsilon; \pi^\varepsilon)$ computed using $\mathcal{M}^\varepsilon$ *(b)* Social utility $U^{\mathrm{S}}(\varepsilon; \pi^\varepsilon)$ in the approval process

Figure 33. **Social utilities under an uncalibrated informative prior.** The left panel shows the social utility (Eq. 13) computed using the belief MDP $\mathcal{M}^\varepsilon$ when the agent uses the optimal policy for each subsidy. The right panel shows the true social utility (Eq. 7) in the approval process when the agent uses the optimal policy π^ε for each subsidy and $\theta^* = 0.65$. The dashed vertical lines correspond to the intervals of the partition $\mathcal{P}$ where the agent's optimal policy is constant (Proposition 8).

Interestingly, Figure 34 shows that the optimal subsidy remains zero as $\rho^{\mathrm{S}}/\rho^{\mathbb{A}}$ increases. At the same time, Figure 35 indicates that our sequential protocol achieves gains exceeding 90% in social utility relative to a non-sequential protocol. In other words, the protocol can substantially improve social utility even in the absence of subsidies.

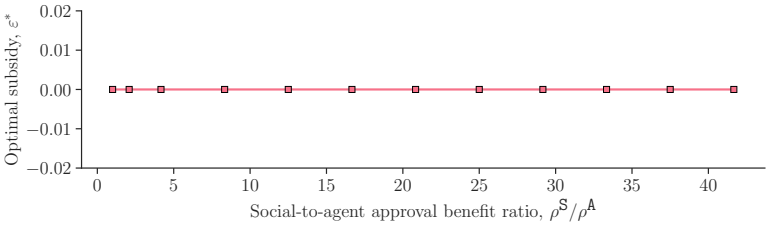


Figure 34. **Optimal subsidy vs. $\rho^{\mathtt{S}}/\rho^{\mathtt{A}}$.** The figure shows, as a function of the social-to-agent approval benefit ratio, the optimal subsidy obtained using Algorithm 1.

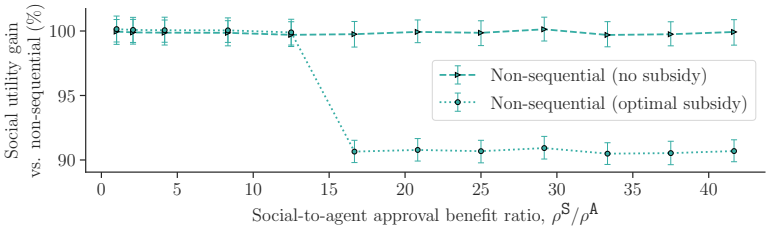
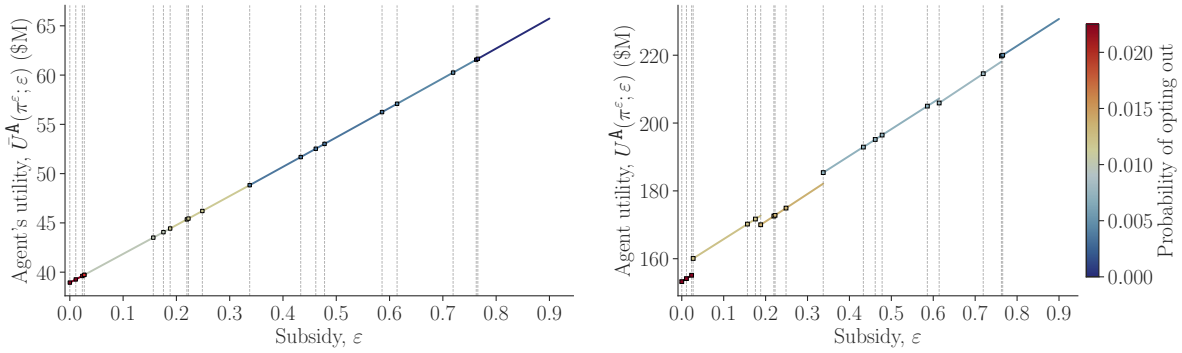


Figure 35. **Social utility gain vs. $\rho^{\mathtt{S}}/\rho^{\mathtt{A}}$.** The figure shows, as a function of the social-to-agent approval benefit ratio, the percentage increase in social utility of the sequential approval protocol relative to a non-sequential approval protocol in which the agent is restricted to a single trial with $n^{\max} = 800$, under (i) the optimal subsidy computed using Algorithm 1 and (ii) no subsidy ($\varepsilon = 0$).

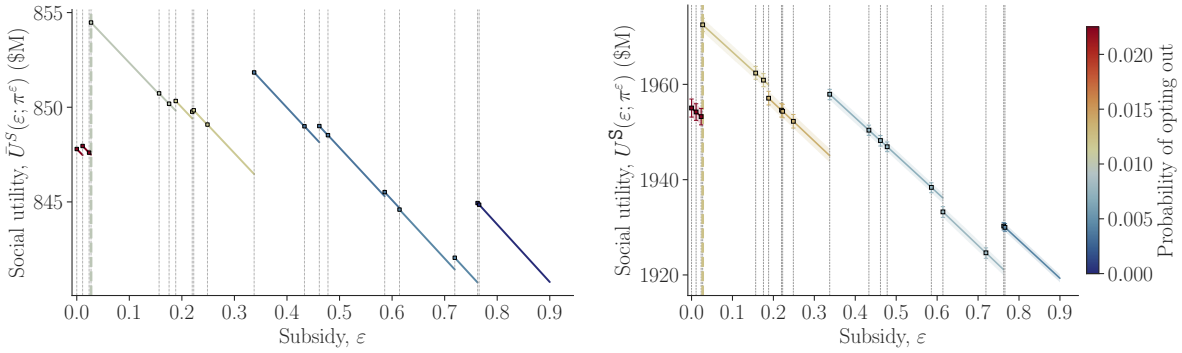
H.2.7. Approval under a different test process

In this section, we present additional experimental results for the antibiotic approval process described in Section 6, using an alternative test process to the process M defined in Proposition 1 and Eq. 3, while keeping all other parameters fixed as in Tables 2 and 3. Here, our goal is to illustrate that Algorithm 1 can provide insight into how to optimally subsidize agents under different statistical tests. More concretely, following Appendix C, we consider the mixed process $M^{\texttt{mix}}$ defined in Eq. 29 with a uniform mixture $P^{\texttt{mix}} = \mathrm{U}(\theta^{\mathrm{b}}, 1)$. We find that the sequential subsidized protocol yields social utility gains of up to 15% compared to a non-sequential baseline without subsidies. Relative to a non-sequential but optimally subsidized baseline, the gains can also reach up to 15% when the ratio $\rho^{\mathrm{S}}/\rho^{\mathrm{A}}$ is low, but diminish rapidly as this ratio increases. The initial optimal action taken by the agent is $n \leq 114$ for all subsidies. Lastly, in Figure 40, we confirm that the mixed process $M^{\texttt{mix}}$ has higher power than M for values of θ^* close to θ^{b}, as discussed in Appendix C.



(a) Agent utility $\bar{U}^{\mathrm{A}}(\pi^{\varepsilon};\varepsilon)$ computed using $\mathcal{M}^{\varepsilon}$ *(b)* Agent utility $U^{\mathrm{A}}(\pi^{\varepsilon};\varepsilon)$ in the approval process

Figure 36. **Agent utilities under a mixed test process.** The left panel shows the agent's utility (Eq. 10) computed using the belief MDP $\mathcal{M}^{\varepsilon}$ when the agent uses the optimal policy for each subsidy, which is a piece-wise linear, convex, and continuous function in accordance with Proposition 8. The right panel shows the true utility of the agent (Eq. 7) in the approval process when using the optimal policy π^{ε} for each subsidy and $\theta^* = 0.65$. The dashed vertical lines correspond to the intervals of the partition $\mathcal{P}$ where the agent's optimal policy is constant (Proposition 8).



(a) Social utility $\bar{U}^{\mathrm{S}}(\varepsilon;\pi^{\varepsilon})$ computed using $\mathcal{M}^{\varepsilon}$ *(b)* Social utility $U^{\mathrm{S}}(\varepsilon;\pi^{\varepsilon})$ in the approval process

Figure 37. **Social utilities under a mixed test process.** The left panel shows the social utility (Eq. 13) computed using the belief MDP $\mathcal{M}^{\varepsilon}$ when the agent uses the optimal policy for each subsidy. The right panel shows the true social utility (Eq. 7) in the approval process when the agent uses the optimal policy π^{ε} for each subsidy and $\theta^* = 0.65$. The dashed vertical lines correspond to the intervals of the partition $\mathcal{P}$ where the agent's optimal policy is constant (Proposition 8).

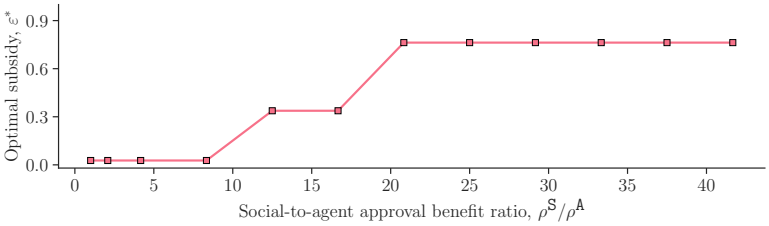


Figure 38. **Optimal subsidy vs. $\rho^{\mathtt{S}}/\rho^{\mathtt{A}}$ using a mixed test process.** The figure shows, as a function of the social-to-agent approval benefit ratio, the optimal subsidy obtained using Algorithm 1.

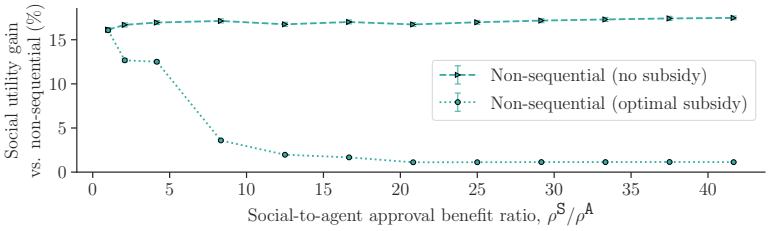


Figure 39. **Social utility gain vs. $\rho^{\mathtt{S}}/\rho^{\mathtt{A}}$ using a mixed test process.** The figure shows, as a function of the social-to-agent approval benefit ratio, the percentage increase in social utility of the sequential approval protocol relative to a non-sequential approval protocol in which the agent is restricted to a single trial with $n^{\mathtt{max}} = 800$, under (i) the optimal subsidy computed using Algorithm 1 and (ii) no subsidy ($\varepsilon = 0$).

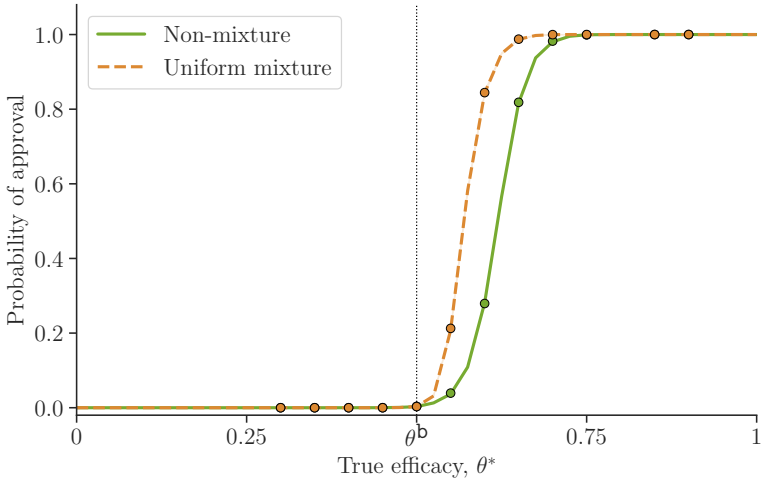


Figure 40. **Probability of approval under the optimal policy and subsidy.** The figure shows, across various efficacies θ^* of the antibiotic, the probability of approval (that is, of rejecting H_0) when the principal selects the optimal subsidy and the agent its optimal policy. The dashed (orange) curve corresponds to the process $M^{\mathtt{mix}}$ defined in Eq. 29 for a uniform mixture (optimal subsidy $\varepsilon^* = 0.027$), while the solid (green) curve corresponds to the process M defined in Proposition 1 and Eq. 3 (optimal subsidy $\varepsilon^* = 0.108$).