Make-or-break: chasing risky goals or settling for safe rewards?

Pantelis P. Analytis[1*], **Charley M. Wu**[2], **Alexandros Gelastopoulos**[3]

[1]Department of Information Science, Cornell University
[2]Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin
[3]Department of Mathematics and Statistics, Boston University

May 2, 2018

Abstract

Humans regularly pursue activities characterized by dramatic success or failure outcomes where, critically, the chances of success depend on the time invested working towards it. How should people allocate time between such *make-or-break* challenges and safe alternatives, where rewards are more predictable (e.g., linear) functions of performance? We present a formal framework for studying time allocation between these two types of activities, and we explore optimal behavior in both one-shot and dynamic versions of the problem. In the one-shot version, we illustrate striking discontinuities in the optimal time allocation policy as we gradually change the parameters of the decision-making problem. In the dynamic version, we formulate the optimal strategy, defined by a *giving-up* threshold, which adaptively dictates when people should abandon the make-or-break goal; we also show that this strategy is computationally unattainable for humans. We then pit this strategy against a boundedly rational alternative using a myopic giving-up threshold that is far simpler to compute, as well as against a simple heuristic that only decides whether or not to start pursuing the goal and never gives up. Comparing strategies across environments, we investigate the cost and behavioral implications of sidestepping the computational burden of full rationality.

Keywords: resource allocation; expectancy theory; dynamic decision making; uncertainty; risky choice; sigmoid curves; bounded rationality.

*Corresponding author. Inquiries related to the paper can be addressed to pantelispa@gmail.com, Department of Information Science, Cornell University, Gates Hall, Ithaca, 14850, NY.

1 Introduction

From applying for a research grant to giving one's all to earn a bonus at work, numerous human activities can either be crowned by dramatic success or thwarted by failure. In these "make-or-break" endeavors, people are either handsomely rewarded upon success or gain nothing upon failure. Although there is almost always a monotonic relationship between the invested time (e.g., of time or effort) in an activity and the expected levels of performance, there is considerable uncertainty about the outcomes (e.g., success or failure). Thus, some of the most important decisions in life may be defined by how one chooses to allocate limited resources between make-or-break tasks—with potentially life changing outcomes—and "safe" alternatives, where outcomes are a more predictable function of performance (e.g., the linear relationship a wage worker experiences between hours worked and income). What is the optimal strategy for allocating time between challenging make-or-break goals and safe alternatives, and how does it compare to simpler, but computationally less expensive, strategies?[1] What behaviors do the different strategies produce across decision-making settings?

As early as the 1930s, Kurt Lewin and his associates identified the importance of success-or-failure reward structures for understanding human behavior (Hoppe, 1931; Lewin, 1936). They investigated how people with different abilities set their aspiration levels (Festinger, 1942; Rotter, 1942), calibrated based on the chances of success and the potential rewards to be gained. Lewin and colleagues postulated that people experience joy when they achieve their aspirations and feel acute loss when they fail, and that they choose the appropriate aspiration level simply by balancing these two motivating forces and the probability of success (Lewin, Dembo, Festinger, & Sears, 1944; Siegel, 1957). In their work, they assumed that people strictly commit themselves to a single aspiration level selected from a continuum of possible choices (i.e., a single task) and that there are no constraints on the amount of time or effort that could be invested (i.e., an infinite horizon). This early work provided valuable insights into how weighing up the chances of success influences the choice of aspiration level. How do these insights transfer to problems where there is scarcity in the time to be allocated among different tasks?[2]

Since then, there have been repeated attempts in the behavioral sciences—independent of or inspired by Lewin's framework—to address how people reason about their chances of success in meeting their goals (e.g., Bandura, 1977; Heider, 1958; Weiner & Kukla, 1970), how aspiration levels affect people's choices (Diecidue & Van De Ven, 2008; J. W. Payne, Laughhunn, & Crum, 1980; Simon, 1955), and how people choose among variable levels of aspiration (Atkinson, 1957). Yet there has been no rigorous mathematical formalization of the problem people grapple with when they have to allocate scarce resources between a make-or-break activity, where investment of additional effort can alter their chances of success, and a safe alternative. The closest attempt to formally ground the problem was a static effort allocation framework

[1] We use time and effort (broadly defined) as equivalent and interchangeable terms. The agent in our story is a human decision maker, but our framework also applies to a firm allocating both human and economic resources towards a project with a success or failure reward structure.

[2] Before the introduction of formal decision theory (von Neumann & Morgenstern, 1944) to psychology by Edwards (1953; 1954), there were several attempts by psychologists to introduce a formal theory of valence (Meehl & MacCorquodale, 1951); many of these attempts were inspired by the study of decision problems characterized by success–failure reward structures. This body of research, commonly referred to as expectancy theory, turned out to be quasi-equivalent to subjective expected utility theory (Feather, 1959; Siegel, 1957), but lacked the theoretical coherence and elegance of the latter, and waned in influence over time.

developed by Kukla (1972), who greatly simplified the problem by assuming away any uncertainty in the relation between effort and performance. Kukla revealed a distinct discontinuous relationship between task difficulty and the optimal allocation of effort; people should refrain from engaging in very difficult tasks, but also invest more effort when success is marginally within reach. Despite its simplicity, the setting Kukla described is of broad interest, because it exposes the non-convex nature of a common class of effort allocation problems encountered in daily life. Similar discontinuities between the parameters that characterize the decision-making problem (i.e., rewards, abilities, constraints) and the normatively prescribed or observed behavior (i.e., time allocation) have been uncovered in various domains, such as allocating time among learning tasks (Son & Sethi, 2006) and allocating attention in perceptual decisions (Morvan & Maloney, 2012), potentially hinting at a common underlying problem structure.

The notion that optimal solutions for even simple problems may harbor non-convexities contrasts with the main time allocation paradigm in economics and psychology, which assumes that different activities, often defined as effort and leisure, complement each other (Becker, 1965; Borjas & Van Ours, 2000; Kurzban, Duckworth, Kable, & Myers, 2013; Varian, 1978). Thus, the resulting allocation problems are convex in nature and easy to solve using standard convex optimization techniques (Boyd & Vandenberghe, 2004). In a similar vein, in optimal foraging theory (the most prominent time allocation paradigm in behavioral biology) the foraging time allocated across different resource patches depends on immediate rewards and the rate at which these rewards diminish (Charnov, 1976). Although these modeling frameworks have been used productively to describe behavior and prescribe how to allocate effort across a wide range of domains, ranging from mental effort allocation (Kool & Botvinick, 2014) to cognitive search (S. J. Payne, Duggan, & Neth, 2007; Pirolli & Card, 1999), they cannot account for the behavioral dynamics observed in decision-making environments with success–failure reward structures. In many real-life problems, people pursue highly rewarding make-or-break goals, where returns depend directly on how additional allocation of time translates into an improved chance of success in the task.

The first contribution of this article is to develop a formal framework that accounts for the role of uncertainty and information in resource allocation problems involving a make-or-break activity and a safe alternative. Our framework draws inspiration from Lewin's early work on aspiration levels (Lewin et al., 1944) and Kukla's (1972) static model of effort allocation, but with a radical departure in how uncertainty accumulates with larger investments of effort or time. We show that manipulating critical factors in the task, such as the rewards or constraints, can produce abrupt and discontinuous changes in the normative allocation strategy, with uncertainty playing a crucial role in determining the optimal solution. We describe two variants of the problem. In the one-shot allocation problem, the decision maker receives feedback about performance only at the very end of the task, after all available time has been allocated. In this formulation of the problem, performance is unobservable during the allocation phase, as is the case when studying for a pass/fail exam or preparing a grant application. In the dynamic allocation problem, the decision maker receives immediate feedback on their current performance and—like a manager trying to meet a performance quota required for a bonus—can dynamically adapt their allocation strategies at any point in time, either continuing to pursue the make-or-break goal or dropping out to allocate time to the safe alternative instead.

The second contribution of this article is to introduce the optimal solution for the dynamic version of

the problem. In line with work in dynamic decision making in management science, economics, finance and mathematical psychology (Dixit & Pindyck, 1994; Malhotra, Leslie, Ludwig, & Bogacz, 2017; McCardle, Tsetlin, & Winkler, 2016; Ulu & Smith, 2009) we show that the solution can be expressed by optimal decision thresholds. We mathematically prove that optimally solving the dynamic allocation problem implies prioritizing investment in the make-or-break task over the safe rewards task (whenever investing in the make-or-break task is considered profitable) and switching unidirectionally to the safe alternative once performance falls below a performance threshold or when the success threshold has been reached. This allows us to use well-known results from optimal stopping theory in stochastic processes (Shiryaev, 2007) as well as standard numerical methods to calculate the optimal giving-up strategy (Kushner & Dupuis, 2013). Following previous work on optimal stopping in economics, finance, and management (e.g., Dixit & Pindyck, 1994), we discretize time and derive the solution by backwards induction. We show that acting optimally implies using a more tolerant giving-up threshold as uncertainty in the environment increases; this finding echoes results from the dynamic investing literature in economics (McDonald & Siegel, 1986). However, the computational burden of the optimal solution puts it well beyond the reach of human decision makers, raising the question of what kinds of cognitively plausible strategies are available to laypeople and managers faced with such tasks.

The third contribution of this article is to analyze how different boundedly rational strategies perform relative to the optimal solution across decision-making environments. First, we define a *myopic giving-up* strategy, which is based on the optimal solution to the one-shot allocation problem. This myopic solution implies giving up earlier than the fully optimal strategy, because it does not consider the possibility of dynamically reassessing one's policy based on new information acquired during the task (i.e., direct feedback about performance). Myopic giving up becomes more conservative relative to the optimal strategy as uncertainty increases, yielding a pattern of behavior similar to that described by risk-averse utility preferences (Pratt, 1964), although here, it is the product of computational limitations. Second, we examine the *play-to-win* heuristic, a simple heuristic strategy that only decides whether or not to invest in the make-or-break task, and then either abandons the task entirely or stubbornly perseveres until a success or failure occurs. When contrasted with optimal giving-up, this strategy produces risk-seeking behavior that is consistent with the sunk cost fallacy (Arkes & Blumer, 1985). Holding all other factors constant, we find that an increase of uncertainty in the environment improves the relative performance of the play-to-win heuristic which, although staggeringly simple, can approximate the performance of the optimal solution. Further, a version of the play-to-win heuristic that relies on a myopic calculation to decide whether to pursue the make-or-break task almost always outperforms the myopic giving-up strategy in terms of expected rewards, even though it disregards most information.

We begin by developing the mathematical framework for the allocation problem and describing how different specifications of the decision-making environment can change the reward structure of the make-or-break task. We then describe the optimal solution to the one-shot allocation problem before moving on to the dynamic version of the allocation problem, where we present both optimal and myopic solutions, and we contrast them with the much simpler play-to-win heuristic. We use expected value theory to build our framework, as it provides the most comprehensible framework to communicate our results. In the Discus-

sion, we draw connections with the literatures on dynamic decision making in mathematical psychology, management science and economics, with expected and non-expected utility theories, and with theories of effort allocation, motivation, and learning. These links highlight a rich set of opportunities for further research and contributions.

2 A Formal Framework

Let us consider an agent who can allocate time (or any another resource) t_m and t_s between two different reward-generating activities $\{m,s\}$, where m is an instance of a *make-or-break* task with a binary success-failure outcome, and s is an instance of a *safe* task, where rewards are a more predictable function of the allocated time. We assume that the total available time to be allocated T is fixed, with $T = t_m + t_s$. We introduce a function describing how invested time maps onto *performance quality* (Equation 1) and a function mapping performance quality onto *rewards* for each reward-generating activity (Equation 2 and 3).

Intuitively, the quality (or quantity) of performance $q(t)$ in any given task is a continuous function of the allocated time t and depends on the abilities of the agent, starting from $q(0) = 0$. It is reasonable to assume that the change in performance Δq over an interval of time Δt is a Gaussian random variable, independent of past performance, and that this random change Δq depends not on when the interval begins but on its length Δt.[3] The unique function that satisfies all of the above assumptions is a diffusion process given by:

$$q(t) = \lambda t + \sigma W_t \tag{1}$$

The first term in Equation 1 is a deterministic term that is proportional to time. We call the coefficient λ the skill level of the agent; larger values of λ lead to better performance in the same amount of time. The second term is a random term, involving a parameter σ and the Wiener process W_t (also called Brownian motion). A Wiener process is the analog of a random walk for continuous time. For each t, it yields a Gaussian random variable $W_t \sim \mathcal{N}(0,t)$, with a mean of 0 and variance equal to t. In the same way that the variance of a random walk increases with the number of steps taken, the variance of a Wiener process also increases with time. The parameter σ is the variance per unit of time; we call it the (performance) uncertainty. As shown in Figure 1, the expectation of performance quality (dotted line) is affected only by the skill level λ and the time t, whereas the underlying uncertainty (confidence band) is determined by the performance uncertainty σ and grows with larger investments of time t.

2.1 Rewards as a Function of Performance

Given a specific reward-generating activity (e.g., writing a grant application or working for an hourly wage), we can describe a function mapping performance onto rewards. Here, we consider two different types of functions, $g_m(x)$ for the make-or-break task and $g_s(x)$ for the safe task, where $x = q(t)$ is the performance quality described previously in Equation 1.

[3] This property is called time homogeneity.

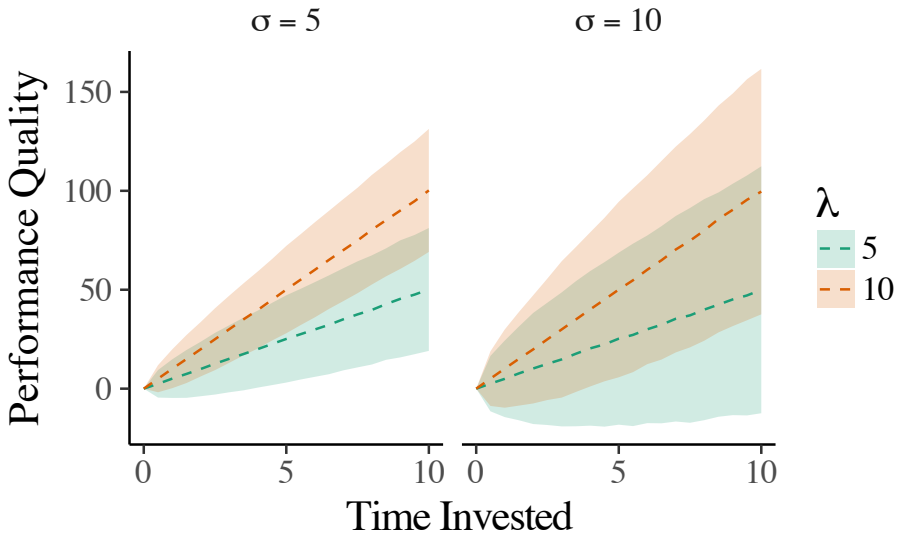

Figure 1: Performance quality as a function of time invested, with total available time $T = 10$. The panels show different levels of performance uncertainty σ; the colors represent different skill levels λ. Dotted lines indicate the expected performance quality and confidence bands show one standard deviation.

2.1.1 Make-or-break.

In the *make-or-break* setting, the agent either receives a considerable reward B upon success, or nothing upon failure, depending on whether their performance $x = q(t)$ reaches a precise success threshold Δ:

$$g_m(x) = \begin{cases} B, & \text{if } x \geq \Delta \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

The performance threshold Δ captures the strict binary outcomes of success or failure and controls the difficulty of the make-or-break task. The reward B may consist, for example, in receiving a monetary bonus at work, winning an award, receiving a grant—or in the joy of achieving an aspiration.[4]

2.1.2 Safe reward.

In the *safe-reward* setting, rewards are a linear function of performance quality, with reward rate v corresponding to an hourly wage or a fixed rate of reward as a function of performance quality $x = q(t)$:

$$g_s(x) = v \cdot x \tag{3}$$

[4]Lewin et al. (1944) and Atkinson (1957) suggest that failures to reach a previously set aspiration level are accompanied by negative payoffs, as people feel embarrassed about the outcome. Similarly, expected utility theory would suggest that people may discount rewards B. Even if we allow for such subjective transformations of the experienced outcomes, the main insights of our article hold.

2.2 Rewards as a Function of Resource Allocation

We can now describe the stochastic processes of earning rewards as functions mapping time allocations t_m and t_s onto rewards r_m and r_s:

$$\begin{aligned} r_m(t_m) = g_m(q_m(t_m)) &= g_m(\lambda_m t_m + \sigma_m W^m_{t_m}) \\ &= \begin{cases} B, & \text{if } \lambda_m t_m + \sigma_m W^m_{t_m} \geq \Delta \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{4}$$

$$r_s(t_s) = g_s(q_s(t_s)) = v \cdot q_s(t_s) = v \cdot (\lambda_s t_s + \sigma_s W^s_{t_s}) \tag{5}$$

The expectation of rewards for the make-or-break task can be written as:

$$\begin{aligned} \mathbb{E}[r_m(t_m)] &= B \cdot \mathbb{P}(\lambda_m t_m + \sigma_m W^m_{t_m} \geq \Delta) \\ &= B \cdot \mathbb{P}\left(\frac{1}{\sqrt{t_m}} W^m_{t_m} \geq \frac{\Delta - \lambda_m t_m}{\sigma_m \sqrt{t_m}}\right) \\ &= B \cdot \left(1 - \Phi\left(\frac{\Delta - \lambda_m t_m}{\sigma_m \sqrt{t_m}}\right)\right) \end{aligned} \tag{6}$$

where Φ denotes the cumulative density function (CDF) of a standard normal distribution. Regardless of the exact parameters of the problem, the above equation is always a sigmoid function (proof provided in Section 6.1 of the Supplementary Material). In contrast, the expectation of rewards for the safe alternative is:

$$\mathbb{E}[r_s(t_s)] = v \cdot \lambda_s t_s \tag{7}$$

Although uncertainty is a crucial component in calculating expected reward for the make-or-break task (Equation 6), the same does not apply to the safe-reward task (Equation 7), because the expected value of the Wiener process $W^s_{t_s} \sim \mathcal{N}(0, t_s)$ is always zero. Thus, the performance uncertainty σ_s does not influence the *expectation*, but only the variance of the reward. To give a better idea of these reward functions, Figure 2 shows the influence of each parameter on expected reward for each activity type as a function of time allocation.

2.2.1 Sensitivity of expected rewards to environmental parameters.

The expected rewards for the make-or-break task are a sigmoid function of time allocation (left column of Fig. 2), where the inflection point and the shape of the expected returns curve are jointly determined by the performance uncertainty σ, the skill level of the agent λ_m, the success threshold Δ, and the bonus B. First, note that the agent's skill level λ_m and the difficulty of the task, as controlled by the success threshold Δ, are directly related; increasing the former has a similar effect to decreasing the latter. Additionally, the performance uncertainty σ determines the shape of the expected returns in the make-or-break task where, in the degenerate case of $\sigma = 0$, expected reward becomes a step function of time allocation (middle left

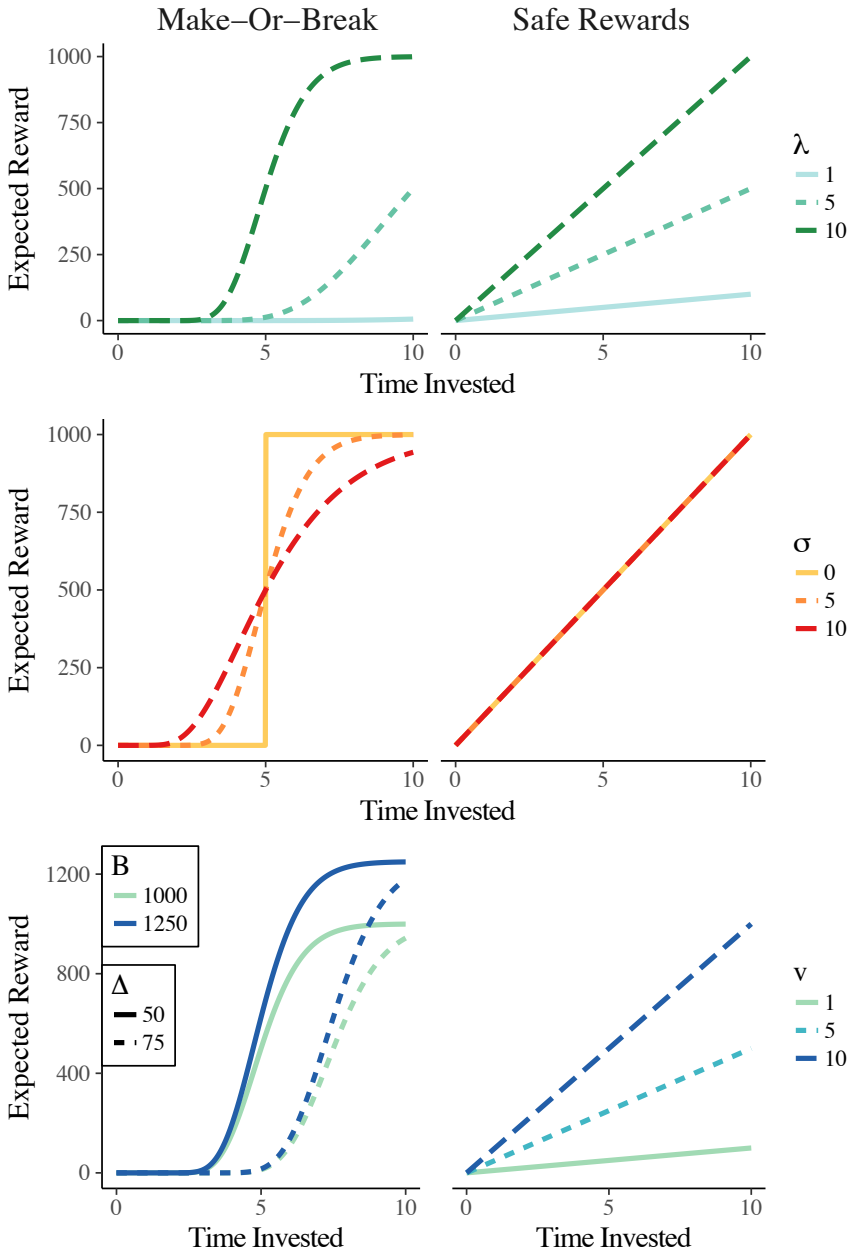

Figure 2: Expected reward as a function of time invested in the make-or-break task (left column) and the safe-reward task (right column). The top row shows the influence of different skill levels (λ) and the middle row shows different levels of performance variance (σ). The bottom row shows the influence of the task-specific parameters: the bonus (B) and success threshold (Δ) bottom left, and the reward rate (v) bottom right. When not specified, we use the default parameter values of $T = 10$, $\lambda_m = \lambda_s = 10$, $\sigma_m = \sigma_s = 5$, $B = 1000$, $\Delta = 50$, and $v = 10$. These parameters are chosen such that the make-or-break reward is equal to the expected reward for the safe-reward task when all available time T is invested, that is, $B = v \cdot \lambda_s T$.

panel of Fig. 2). As uncertainty increases, the curvature of the sigmoid reward function becomes smoother. Finally, the bonus B in the make-or-break task and the skill level λ_s and reward rate v in the safe-reward task control the relative payoffs between the two tasks. Multiplying or dividing both B and $v \cdot \lambda_s$ by the same amount does not alter the relative payoff in the tasks.

2.2.2 Marginal returns.

Investing a small amount of time in the make-or-break task typically yields a very small expected return; the resulting performance is almost certain to fall below the success threshold. Initially, an increasing time allocation in the make-or-break task corresponds to only small increases in expected reward, with returns being smaller when uncertainty is low. However, as one continues to invest more time in the make-or-break task, the rate of increase in expected reward begins to accelerate and rises rapidly until it reaches an inflection point (at $t_m = 5$, middle left panel of Fig. 2). The rate at which expected reward increases around the inflection point is determined by performance uncertainty σ, with higher uncertainty creating a more gradual transition (smoother sigmoid curves; see middle left panel of Fig. 2). Eventually, as performance quality surpasses the threshold Δ, the marginal increases in expected reward become smaller and smaller, and the expected reward curve draws near the upper bound of B. Thus, after a certain expected performance level, additional allocation of time has diminishing marginal returns.[5]

3 Results

Having established the formal framework for the problem, we now present solutions to two versions of the problem and discuss their implications. First, we address the *one-shot* version of the problem, where the agent makes a single decision about how to allocate resources between the two tasks, without any feedback on performance or reward. Second, we address the *dynamic* version of the problem, where performance $q(t)$ is observable and switching between activities can occur at any point in time $t < T$ (i.e., after any amount of investment prior to exhausting all available time). We present a fully rational and a boundedly rational solution to the dynamic problem, both relying on the idea of a "giving-up" threshold; we then compare these solutions with a simple heuristic strategy that assumes that agents stubbornly pursue success in the make-or-break task and never give up.

3.1 The One-Shot Allocation Problem

If an agent has a fixed amount of time T to allocate between the two tasks, and needs to commit to this allocation at the beginning of the task, how can they maximize the sum of expected rewards? This problem is identical to a case where the agent has to commit to a plan upfront, or to a scenario where there is no feedback on performance. Therefore, the optimization problem is reduced to a single decision about how to

[5] Heath, Larrick, and Wu (1999) have postulated a sigmoid function governing the evaluation of people's performance in relation to precise goals. In their framework, they employ prospect theory (Kahneman & Tversky, 1979) to devise the subjective reward function, setting the goal as a reference point and assuming away uncertainty about performance. In our model, in contrast, the shape of the sigmoid function is produced due to uncertainty.

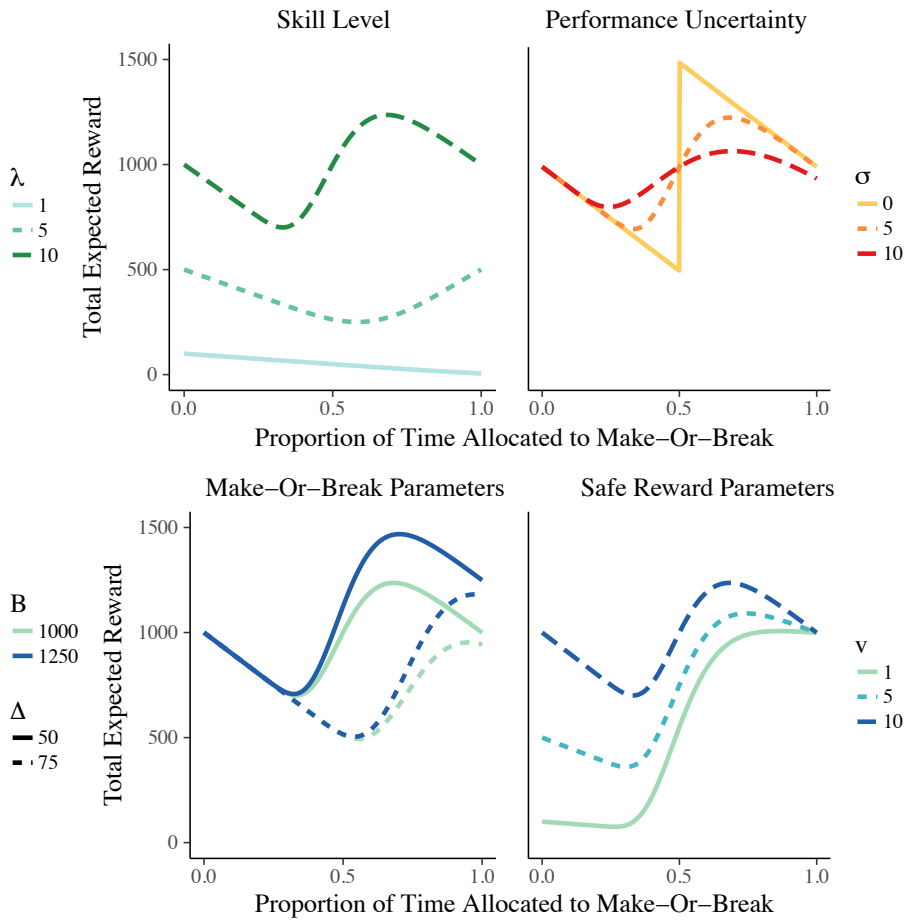

Figure 3: Total expected reward as a function of the proportion of time allocated to the make-or-break task, where remaining time is allocated to the safe-reward alternative. The top left panel shows how the shape of the curve varies for different skill levels (we set $\lambda_m = \lambda_s = \lambda$). The top right panel shows the effect of uncertainty (in the make-or-break task, $\sigma = \sigma_m$) on the curve shape. The bottom panels show how the returns curve changes as we vary the bonus or the threshold (left) or the reward rate in the safe task (right). When not specified, we use the following default parameter values: $T = 10$, $\lambda_m = \lambda_s = 10$, $\sigma = 5$, $B = 1000$, $\Delta = 50$, $v = 10$.

divide the total available time T between the two activities t_m and t_s, where the overall reward function can be written as:

$$\begin{aligned} h(t_m, t_s) &= \mathbb{E}[r_m(t_m)] + \mathbb{E}[r_s(t_s)] \\ &= B\left(1 - \Phi\left(\frac{\Delta - \lambda_m t_m}{\sigma_m \sqrt{t_m}}\right)\right) + v \cdot \lambda_s t_s. \end{aligned} \tag{8}$$

An agent should then optimize this function (Equation 8) under resource constraints, which is written formally as:

$$\begin{aligned} \underset{t_m, t_s}{\text{maximize}} \quad & h(t_m, t_s) \\ \text{subject to} \quad & T = t_m + t_s \end{aligned} \tag{9}$$

The above equations give rise to a simple, yet commonly encountered, non-convex optimization problem. Below, we describe and offer intuitions about how the optimal allocation policy depends on the parameters of the problem (see Fig. 3).

3.1.1 Playing safe or going all in.

For some sets of parameters, one option completely dominates the other, and the decision maker should simply invest all available time resources T in just one of the tasks. In some cases, for instance, the rewards of the make-or-break task simply do not justify the risks, and the clear dominant option is to "play it safe" by investing all available time resources in the safe-reward task. This is the case when (i) there is only a slim chance of success in the make-or-break task (i.e., due to scarcity of total available time T, a very high threshold Δ, or low skill level λ_m; see top left of Fig. 3 where $\lambda_m = 1$) or (ii) when the expected returns from the safe-reward task are simply much higher (i.e., when $v \cdot \lambda_s T$ is much larger than B).

As we vary the parameters making the make-or-break task more attractive, it becomes rational for agents to allocate at least some—possibly all—of their time to the make-or-break task. There is a discontinuous switch in the optimal policy when a critical point in the parameter space is crossed as we increase the total available time, the bonus of the make-or-break task, or the skill level of the agent (see Section 7.2 in the Supplementary Materials). This is illustrated in Fig. 4, where the optimal policy switches from "playing it safe" ($t_s = T$) to "going all in" ($t_m = T$) at the critical value of λ_m satisfying the equation

$$B \cdot \mathbb{P}(\lambda_m T + \sigma_m W_T^m > \Delta) = v \cdot \lambda_s T \tag{10}$$

Equation 10 corresponds to the point where the two extreme policies of "playing it safe" and "going all in" produce equal expected rewards. Transitions from one extreme to the other always occur when the make-or-break task is sufficiently rewarding (see Supplementary Materials for proof), and we vary the total available time resources T. In the Supplementary Material we also specify the exact conditions under which all-or-nothing transitions should occur when we vary the skill level or the reward of the make-or-break task.

3.1.2 Slacking off for the highly skilled.

In environments where success in the make-or-break task is highly likely without requiring the investment of all available time, the optimal policy entails allocation of resources to both activities. As expected performance in the make-or-break task increases beyond the critical point described in Equation 10, the decision maker can potentially err on the side of investing too much time in the make-or-break task. This error comes with the opportunity cost of losing potential rewards from investing surplus time in the safe-reward task. This will be the case if the marginal expected gain of investing more time in the make-or-break task is less

than the gain of investing more time in the safe-reward task, that is:

$$[\mathbb{E}[r_m(t_m)]]'_{t_m=T} < v \cdot \lambda_s, \tag{11}$$

where the derivative is taken with respect to the allocation t_m. In Figure 4, this corresponds to the point at which the optimal allocation of time to the make-or-break task begins to decrease, as the likelihood of success increases as a function of skill level (λ_m). Everything else being equal, a lower threshold and more total available resources make it more likely that an agent should allocate time to both activities. People with higher skill levels can quickly secure a higher probability of success and invest residual time resources in the safe-reward task. Whereas Kukla (1972) suggested a discontinuous relationship between the difficulty of a task and the effort people exert here we have generalized it to any level of uncertainty and clarified the relations between the different parameters of the problem (see Fig 4, where skill and difficulty are inverse concepts).

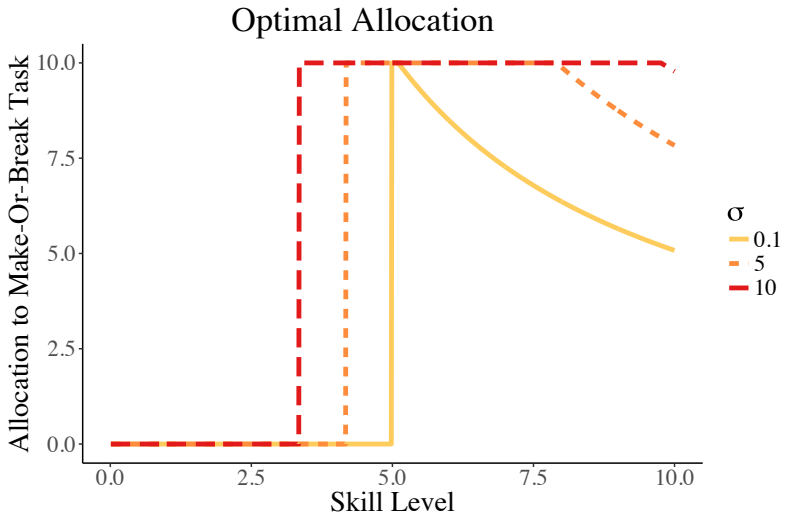

Figure 4: Optimal allocation as a function of skill level and under different levels of performance uncertainty (line type). Unless otherwise specified, we use the following default parameter values: $T = 10$, $B = 1000$, $\Delta = 50$, $v = 10$. To illustrate the crucial role of uncertainty, we set $\lambda_s = 3$ and vary the value of $\lambda_m \in [0, 10]$. For this set of parameters, it is already profitable to switch from allocating everything to the safe-reward task to allocating everything to the make-or-break task when an agent has a 30% chance of success. At higher uncertainty levels, this critical point occurs at an even lower level of skill (compare the optimal allocation policy for $\sigma = 0.1$ with $\sigma = 5$ and $\sigma = 10$). On the other hand, people with higher skill levels should allocate time to the make-or-break task until the point that marginal returns from increasing the chances of success in the task are equal to the linear opportunity cost of the safe-reward task (see Equation 11). At higher uncertainty levels, this critical point occurs at higher skill levels.

3.1.3 The motivating effect of uncertainty.

Everything else being equal, higher performance uncertainty implies that it is optimal to allocate all time to the make-or-break task for a larger subset of parameters in the parametric space. In everyday language, this means that for higher levels of uncertainty, people with lower skill should dedicate themselves completely to the make-or-break goal in hope of eventual success. It also makes sense for people with higher skill to put all their effort in the make-or-break task, in order to minimize the risk of failing (see also Figure

8 in the Supplementary Material). Higher uncertainty leads to an increase in the expected rewards of the make-or-break task for small amounts of time allocation, and it also implies that the deceleration of marginal returns occurs more slowly (see the middle left panel of Figure 2 and the upper right panel of Figure 3). As a result, in more uncertain environments (i.e., larger σ), the critical point described by Equation 10 occurs for lower levels of skill or total available time, while the critical point described by Equation 11 occurs for larger values. Figure 4 demonstrates this effect when comparing a low uncertainty environment ($\sigma = 0.1$) with intermediate ($\sigma = 5$) or high uncertainty environments ($\sigma = 10$). The motivating effect of uncertainty is particularly pronounced when comparatively large rewards are at stake in the make-or-break task (i.e., a high $B/(v \cdot \lambda_s T)$ ratio) and gradually attenuates as the relative rewards from the safe-reward task increase.

3.1.4 Discussion.

In everyday experience, people often need to decide how to allocate their time or effort between challenging make-or-break goals and safe alternatives. We have shown that this type of resource allocation problem is non-convex in nature, leading to striking discontinuities in how people should allocate their time as the parameters of the problem vary. Small differences in people's skill levels, resources, or the rewards available may change the optimal policy from allocating all of one's time to the safe-reward task to investing everything in the make-or-break task. Crucially, we have shown how the degree of uncertainty in the environment moderates the optimal allocation policy and determines when these discontinuities are expected to occur.

3.2 The Dynamic Allocation Problem

So far we have assumed that the agent makes a single allocation decision, dividing the available time between two activities without any feedback on performance. In many problems, however, people dynamically obtain information about their performance, which they can use to reassess their prospects of success and alter their behavior on the fly. How should agents dynamically invest their time when they can directly observe their performance and switch dynamically between tasks?

3.2.1 Optimal allocation policy.

Order of tasks and switching between tasks. Intuitively, people above a certain skill level in the make-or-break task should start by investing time in it; they can switch to the safe-reward task in case of an early success or if success seems unattainable. Because future rewards in the safe-reward task do not depend on past performance (i.e., there is no useful feedback from the safe task), the dominant strategy is to begin with the make-or-break task and to switch unidirectionally to the safe-reward task (see Section 6.3 in the Supplementary Material for a proof). This leads to a significant simplification of our problem: an optimal solution to the allocation problem involves at most one switch between tasks and only from the make-or-break to the safe-reward task. The question then is when to make this single switch.

Giving-up threshold. At any given point in the make-or-break task, the agent has to decide whether to switch immediately to the safe-reward task or to further pursue the make-or-break goal, based solely on their

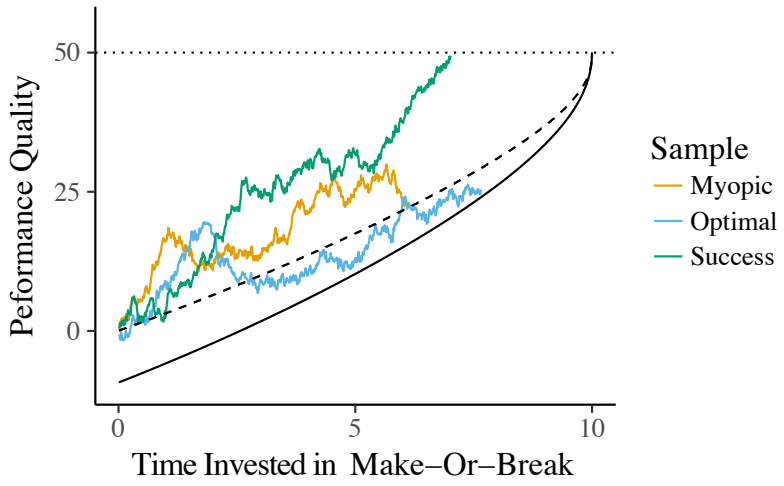

Figure 5: Visualization of the *myopic* (dashed line) and *optimal* (solid line) giving-up thresholds for three samples drawn from the stochastic performance process $q(t)$. The green sample reaches the success threshold (dotted line) and is free to invest the remaining time in the safe-reward task. The orange sample has hit the myopic giving-up threshold (dashed line) and therefore gives up on the make-or-break task. However, an optimal decision maker has a more tolerant giving-up threshold (solid line), thus the blue sample and the green sample continue even after hitting the myopic threshold, resulting in either a later giving-up point (blue) or eventually reaching the success threshold (green). Any time resources remaining after the agent reaches either a giving-up threshold or the success threshold are allocated to the safe-reward task. We use the following default parameter values: $T = 10$, $\sigma_m = \sigma_s = 5$, $B = 1000$, $\Delta = 50$, $v = 10$, but use $\lambda_m = \lambda_s = 5$ for illustrative purposes.

current performance. At very poor performance levels, success is highly unlikely; therefore, it makes sense to switch to the safe-reward task. On the other hand, if the agent is relatively close to the success threshold (and assuming the reward is large enough), it makes sense to continue in the hope of an eventual success. For a specific intermediate level of performance, the two strategies will yield the same expected reward, such that the agent can be indifferent about which strategy to follow. We call this indifference point the *giving-up threshold* and at time t we denote it by c_t. Intuitively, the agent should abandon the make-or-break task in favor of the safe-reward task if, and only if, performance is equal to or falls below this value after time t. Assuming that the agent starts with a performance above the giving-up threshold (i.e., $q(t_0) > c_0$), the agent should give up whenever performance $q(t_m)$ falls below the giving-up threshold c_t or upon reaching the success threshold Δ. More precisely, the point at which an agent should give up is given by

$$\tau = \min\{t : q_m(t) \leq c_t \text{ or } q_m(t) \geq \Delta\}. \tag{12}$$

In the stochastic processes literature, this type of problem is considered a problem of *optimal stopping*. The existence of a giving-up threshold that leads to an optimal policy through Equation 12 is guaranteed by Theorem 2.2 in Peskir and Shiryaev (2006); several methods to find this optimal policy are given in Chapter 8 of the same book.[6] To the best of our knowledge, there is no analytic solution for finding the

[6]There are various types of optimal stopping problems (for examples, see DeGroot, n.d.). The dynamic decision making problem we study shares concepts and methods with three distinct lineages of such problems. The first can be traced to Wald's sequential analysis (1945), where an agent has to decide when to stop evaluating alternative hypotheses. This line of thought has been pursued further in psychology and neuroscience, leading to choice models where evidence in favor of different alternatives unfolds dynamically in time as a stochastic process (Ratcliff, Smith, Brown, & McKoon, 2016; Tajima, Drugowitsch, & Pouget, 2016). A

optimal policy and all approaches that can be used to find it are computationally expensive. Here, we follow the most widely used approach for deriving the optimal policy, which involves discretizing time and using backwards induction (Dixit & Pindyck, 1994; Kushner & Dupuis, 2013).

First, we discretize time by dividing T into n equally sized intervals. This turns the problem into a discrete optimal control problem whose solution involves solving the resulting Bellman Equation using backwards induction

$$R_k(x) = \max\left\{\int_{-\infty}^{\infty} R_{k+1}(y)\cdot\phi_x(y)dy,\ v\cdot\lambda_s\cdot\frac{(n-k)T}{n}\right\}, \tag{13}$$

for $k = 0,\ldots,n$. $R_k(x)$ is the expected total reward from time k to n, if at k the performance is x, and ϕ_x is a suitable normal distribution. The initial conditions are given by $R_n(y) = r_m(y)$ (see Section 6.4 in the Supplementary Material for details). Thus, the giving-up threshold c_k at time k is the unique solution of the equation

$$\int_{-\infty}^{\infty} R_{k+1}(y)\cdot\phi_{c_k}(y)dy = v\cdot\lambda_s\cdot\frac{(n-k)T}{n}. \tag{14}$$

The optimal policy is to switch to the safe-reward task at the first step k such that $q_m(t_k) \leq c_k$, due to falling below the threshold, or $q_m(t_k) \geq \Delta$, due to success. Deriving the optimal threshold requires starting from the end of the task and reasoning backwards. Thus, calculating the giving-up threshold at t_0 requires knowledge of the threshold values at all subsequent steps. Finding the optimal solution is computationally demanding not only for humans but also for machines, as we explain in the following paragraph.

The computational burden of calculating the optional threshold. Herbert Simon, one of the founders of artificial intelligence, argued that theories of rationality that do not take into account the computational costs of problem solving are incomplete or even misleading (Simon, 1978). We build on his rich conception of rationality and apply complexity theory (Papadimitriou, 2003; Van Rooij, 2008) to express the computational cost of the optimal stopping strategy. As shown in Equation 13, for each step k and each performance value x, we have to calculate an integral over all performance values y at the next step. Assuming that the integral is computed numerically and approximated by a sum, both its computational cost and the accuracy of the result depend on the number of summands we use. If we use m terms in the sum, then the computational complexity of one integration will be $O(m)$. This will give us $R_k(x)$ for a specific k and x. But in order to later find $R_{k-1}(x')$, for any x', we need to have $R_k(x)$ for m different values of x. Thus, for each k, we need $O(m^2)$ computations. This gives us $R_k(x)$ for a specific k and for m different values of x. The accuracy of the approximation also depends on the number of steps we use in the time axis, that is, the number of values of k. Thus, if we divide time into n steps, we need in total $O(n\cdot m^2)$ computations for the solution described above. Even for a relatively small number of discrete time steps and summands used for integration, these operations are very expensive to compute, regardless of the exact cost measure used to penalize excessive

second lineage studied in economics and finance builds on stochastic processes to investigate when to make high-stakes decisions when the value of assets fluctuate (Dixit & Pindyck, 1994; Jacka, 1991). The main theoretical result of this line of research is that agents should wait longer before making consequential decisions. A third strand of research in management science and operations research relies on dynamic optimization techniques to explore when to adopt a new technology or give up investing in it when new information about its potential unfolds dynamically (McCardle, 1985; Ulu & Smith, 2009).

computation (for different operationalizations of computational costs, see Fechner, Schooler, & Pachur, 2018; Johnson & Payne, 1985; Lieder & Griffiths, 2017). For instance, calculating the optimal policy for $n = m = 10^2$ requires a million computations ($n \cdot m^2 = 10^6$). Thus, the level of computation required for this strategy places it firmly outside the bounds of human rationality and would be non-negligible even for state-of-the-art computational systems.

3.2.2 Boundedly rational alternatives.

Several studies have shown that most people do not use backward induction in dynamic decision-making settings, even in problems with a shallow planning depth (e.g., Hotaling & Busemeyer, 2012; Huys et al., 2015; Zhang & Yu, 2013). The same holds in game-theoretical contexts (Johnson, Camerer, Sen, & Rymon, 2002; McKelvey & Palfrey, 1992). Instead of using computationally complex solutions, people tend to rely on simple and computationally inexpensive algorithms, which in some environments lead consistently to deviations from optimality (Lieder & Griffiths, 2017; Tversky & Kahneman, 1974) but in other environments are surprisingly close to the optimal solution (Gaissmaier & Schooler, 2008; Hey, 1982). In the following sections, we examine two boundedly rational strategies that circumvent the costs of full rationality.

The myopic giving-up strategy. An alternative to backward induction is to assume that the decision maker acts myopically and decides whether or not to pursue the make-or-break task further as if this decision were the last that they could make. Myopic strategies accurately describe human behavior in a wide array of settings, ranging from sequential hypothesis testing to sequential search and multi-armed bandit tasks (see also Busemeyer & Rapoport, 1988; Gabaix, Laibson, Moloche, & Weinberg, 2006; Stojic, Analytis, & Speekenbrink, 2015; Wu, Schulz, Speekenbrink, Nelson, & Meder, 2018; Zhang & Yu, 2013). In our case, as more time is allocated to the make-or-break task, some of the associated uncertainty becomes replaced by an actual outcome $y = q_m(t')$ experienced up until time $t' \in [0, T]$. The agent can use this information to revise their allocation policy, making a myopic decision at any point during the allocation problem. Because future performance is independent of past performance (by the properties of the Wiener process), the problem of finding the optimal allocation is the same as before, but with total time $T_r = T - t'$ and threshold $\Delta - y$. The expected reward for the make-or-break task, if additional time t_m were invested in it, would be

$$\begin{aligned} \mathbb{E}[r_m(t_m)] &= B \cdot \mathbb{P}(\lambda_m t_m + \sigma_m W^m_{t_m} \geq \Delta - y) \\ &= B \cdot \left(1 - \Phi\left(\frac{\Delta - y - \lambda_m t_m}{\sigma_m \sqrt{t_m}}\right)\right) \end{aligned} \tag{15}$$

The expected returns for the safe-reward option would still be $E[r_s(t_s)] = v \cdot \lambda_s t_s$. A myopic agent seeks to optimize the sum of these rewards, under the constraint $T_r = t_m + t_s$, and will continue to invest in the task for as long as it would myopically be profitable to do so. A giving-up threshold in terms of performance can be computed (Fig. 5, dashed line) as the point below which the myopic policy prescribes investing no further time in the make-or-break task, in other words when $t_m = 0$. When performance drops below this threshold, the myopic agent will switch to the safe-reward task.

Finding the t_m that maximizes the total expected reward is an easy optimization problem, which can be

solved by finding the roots of the derivative of Equation 15. The computational complexity of this, using Newton's method, for example, is a constant multiple of the complexity of computing elementary functions (Brent, 1976). Hence, at any point, an agent can decide whether to continue with the make-or-break task by solving a computationally inexpensive problem. In total, $O(n)$ calculations will be needed, but distributed equally over n intervals. This starkly contrasts with the optimal strategy, where the entire computational cost has to be paid upfront.

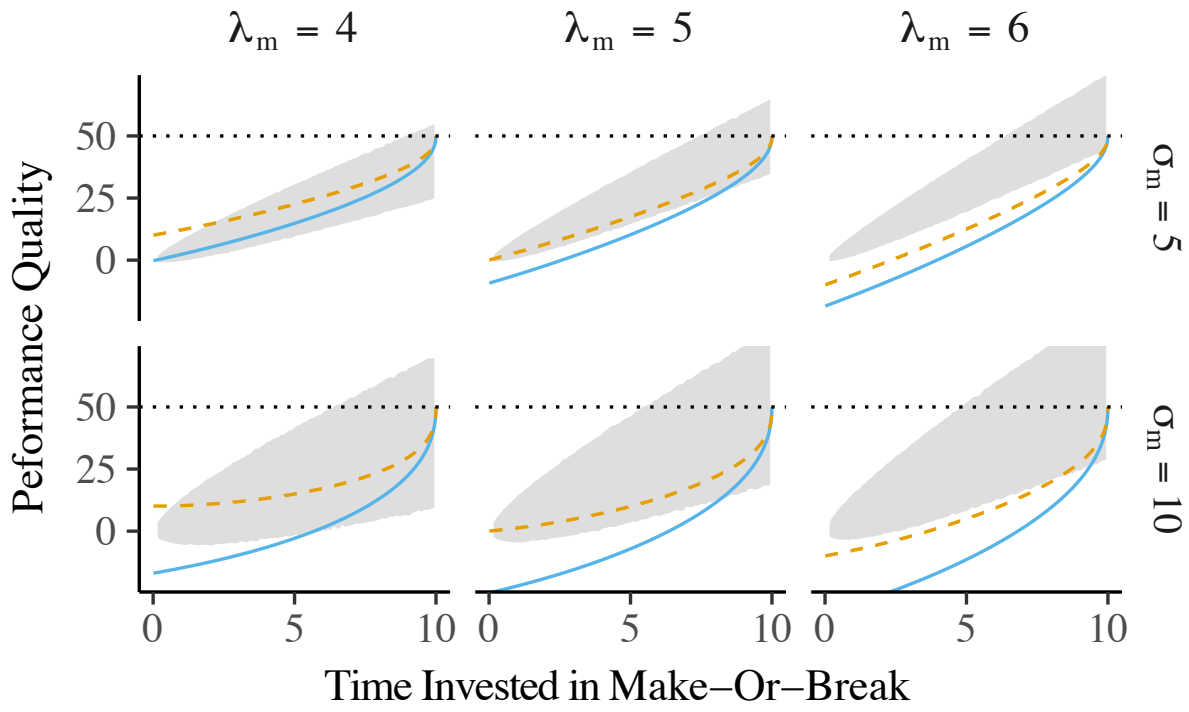

Figure 6: Myopic (dashed line) and optimal (solid line) giving-up thresholds, as a function of performance variance (σ_m) and ability (λ_m) in the make-or-break task. Other parameters are defaulted to $T = 10$, $\sigma_s = 5$, $B = 1000$, $\Delta = 50$, $v = 10$, and $\lambda_s = 5$. The gray confidence bound shows one standard deviation of performance quality. As σ_m increases, the differences between the myopic and optimal thresholds increase. Note that the myopic threshold is more conservative than the optimal threshold.

The play-to-win strategy. Another way to cope with the computational complexity of dynamic decision making is to apply a simple heuristic (Gigerenzer & Gaissmaier, 2011; Tversky & Kahneman, 1974). We consider the play-to-win strategy, a heuristic strategy that stubbornly pursues success in the make-or-break task, switching to the safe-reward task only when the make-or-break threshold is reached. Thus, the strategy only needs to decide whether to start pursuing the make-or-break goal at all, eliminating the computational costs almost entirely. Although the strategy is suboptimal, it has an intuitive appeal. Most people are familiar with real-world examples of individuals who, once they had taken the first step, would not give up on a goal, irrespective of their chances of success. The crucial event calibrating the expected returns on this strategy is the first time point at which the stochastic process crosses the success threshold. This event is commonly referred to as the first passage time of the stochastic process. First passage time problems for

a Wiener process with a single threshold have been studied extensively across scientific disciplines (e.g., Lancaster, 1972; Lee & Whitmore, 2006; Redner, 2001) and the results can be readily transferred to our setting. To calculate the expected total reward for this strategy, we need to bear in mind that there are two cases: either the agent pursues the make-or-break task, but never reaches the reward threshold Δ and receives zero reward, or the agent reaches the reward threshold after some time $\tau \leq T$ and receives the bonus B in addition to rewards from the safe-reward task, to which any surplus resources $T - \tau$ were allocated. The time τ can be defined mathematically as:

$$\tau = \min\{t : q_m(t) \geq \Delta\} = \min\left\{t : W_t^m + \frac{\lambda_m}{\sigma_m} \cdot t \geq \frac{\Delta}{\sigma_m}\right\} \tag{16}$$

In the stochastic processes literature, the above is referred to as a hitting time or first passage time, at level $\frac{\Delta}{\sigma_m}$, of a Wiener process with drift $\frac{\lambda_m}{\sigma_m}$. In Section 6.5 of the Supplementary Material, we provide the exact probability density function of τ (known as an inverse Gaussian distribution) and an analytical expression for the expected reward for this strategy.

What is a good starting rule for the play-to-win strategy? For somebody in possession of a calculator and the analytical formula for the expected returns on the strategy (see Section 6.5 in the Supplementary Material), it would be computationally trivial to check whether the expected returns are higher than for playing it safe. Even if suboptimal, this rule ensures that the agent will reap—whenever possible—the higher expected returns procured from stubbornly pursuing the make-or-break goal. However, it is unlikely that laypeople will be able to use the analytical formula. Alternatively, an individual could employ a myopic calculation (using the myopic strategy only at $t = 0$) to gauge whether to start pursuing the make-or-break goal in the first place, and then never give up. Such starting rules would protect people from starting in contexts where success is unlikely (e.g., where the decision maker has a low skill level and the uncertainty in the environment is relatively low), eliminating the costs of continuously monitoring their progress and calculating a giving-up threshold.

3.2.3 Strategy comparison.

For the play-to-win heuristic, the expected returns and the distribution of the time of success can be derived analytically; for threshold-based strategies, in contrast, the expected returns can be derived only numerically and the distribution of success and dropout times, only via simulations. To obtain additional insights into how the strategies compare to one another, we therefore simulated the behavior of decision makers in the make-or-break task at three skill levels ($l_m = 4, 5, 6$) and two levels of uncertainty ($\sigma_m = 5, 10$). These conditions are sufficient to capture the crucial factors influencing the performance of the strategies as well as the interaction between them. Figure 6 shows the optimal and the myopic giving-up thresholds; Figure 7 compares the simulated behavior of agents applying the two threshold strategies as well as the play-to-win heuristic. Figure 7 illustrates the earnings of the strategies for 10,000 random samples generated from the Wiener process with the corresponding parameters (average earnings represented by diamonds), with each dot corresponding to a single instantiation of the stochastic process.

Threshold divergence and observed behavior. As uncertainty in the environment increases, so does the value of information that will be received through feedback in the course of the task. Only the optimal giving-up strategy takes into account the value of information that will be revealed in the future, and as a result prescribes persevering for longer (before dropping out) than the myopic giving-up strategy does (Fig. 6). An agent following the myopic strategy will start investing in the make-or-break task only if $B \cdot \left(1 - \Phi\left(\frac{\Delta - y - \lambda_m T}{\sigma_m \sqrt{T}}\right)\right) \geq v \cdot \lambda_s T$, otherwise choosing to play it safe. In contrast, an optimal threshold agent will be willing to try their chances even at lower levels of skill λ_m. The two thresholds diverge more as uncertainty increases, with the optimal threshold becoming more and more tolerant relative to the myopic threshold. This result echoes the main finding from optimal stopping models in economics and finance, which suggest that optimally behaving agents should wait longer before making high-impact, irreversible decisions, such as selling a factory or getting married (see Dixit & Pindyck, 1994). Threshold divergence does not, however, always imply that the behaviors of people following the two strategies will deviate. At very low or very high skill levels, the optimal and myopic giving-up strategies lead to similar observed behavior and earned rewards, despite the threshold differences. At very low skill levels, both strategies play it safe; at very high skill levels, even the myopic giving-up threshold—which is more conservative—is placed far below the expected performance, and rarely triggering a drop out.[7]

Optimal vs. myopic giving-up strategies. The behavior of agents relying on the optimal or myopic giving-up strategies depends on (i) the skill levels of the agents and (ii) the amount of uncertainty in the environment. In this section, we illustrate these two effects and their interplay, focusing on levels of skill for which the strategies diverge. Everything else being equal, the results will be more pronounced when a larger reward is at stake (for high $B/(v \cdot \lambda_s T)$ ratios), and the differences will attenuate when the relative rewards from the safe-reward task increase. Here we present results for $B = 2v \cdot \lambda_s T$.

The effect of skill level: The behavior of the two strategies diverges as soon as it becomes optimally profitable to allocate any amount of resource to the make-or-break task. For instance, at relatively low skill levels and intermediate levels of uncertainty (e.g., $\sigma_m = 5, \lambda_m = 4$), an optimal decision maker should start pursuing the make-or-break task, whereas a myopically acting agent will play it safe. Even under these conditions, an unexpectedly good start could bring the success threshold within reach for an optimal decision maker. If performance falls far below expectations, a decision maker can quickly withdraw to the safe-reward alternative, thereby acquiring at least some of the rewards from the safe task. Only a few of the simulated agents following the optimal strategy persevere until success (only 3.3 % for $\sigma_m = 5$ and $\lambda_m = 4$); most simulated agents drop out early (see the upper left panel of Figure 7). Under this set of parameters, the optimal policy leads to a modest increase in terms of expected returns.[8]

The divergence in observed behavior and earned rewards between the myopic and optimal giving-up strategies is even more pronounced at intermediate skill levels (i.e., $\sigma = 5, \lambda_m = 5$). Under these conditions,

[7] Note that for $\sigma = 0$ the problem degenerates and both strategies follow the same solution as the one-shot problem. The notion of a giving-up threshold is practically meaningless in such a scenario, because the agent can decide on the optimal time allocation from the outset, and there is no reduction in uncertainty through observing performance.

[8] Our simulation framework can be used to study the performance of any other strategy relying on a giving-up threshold. Note that the play-to-win heuristic can also be seen as a boundary case of a strategy with a giving-up threshold, where the threshold value is set to negative infinity at all times.

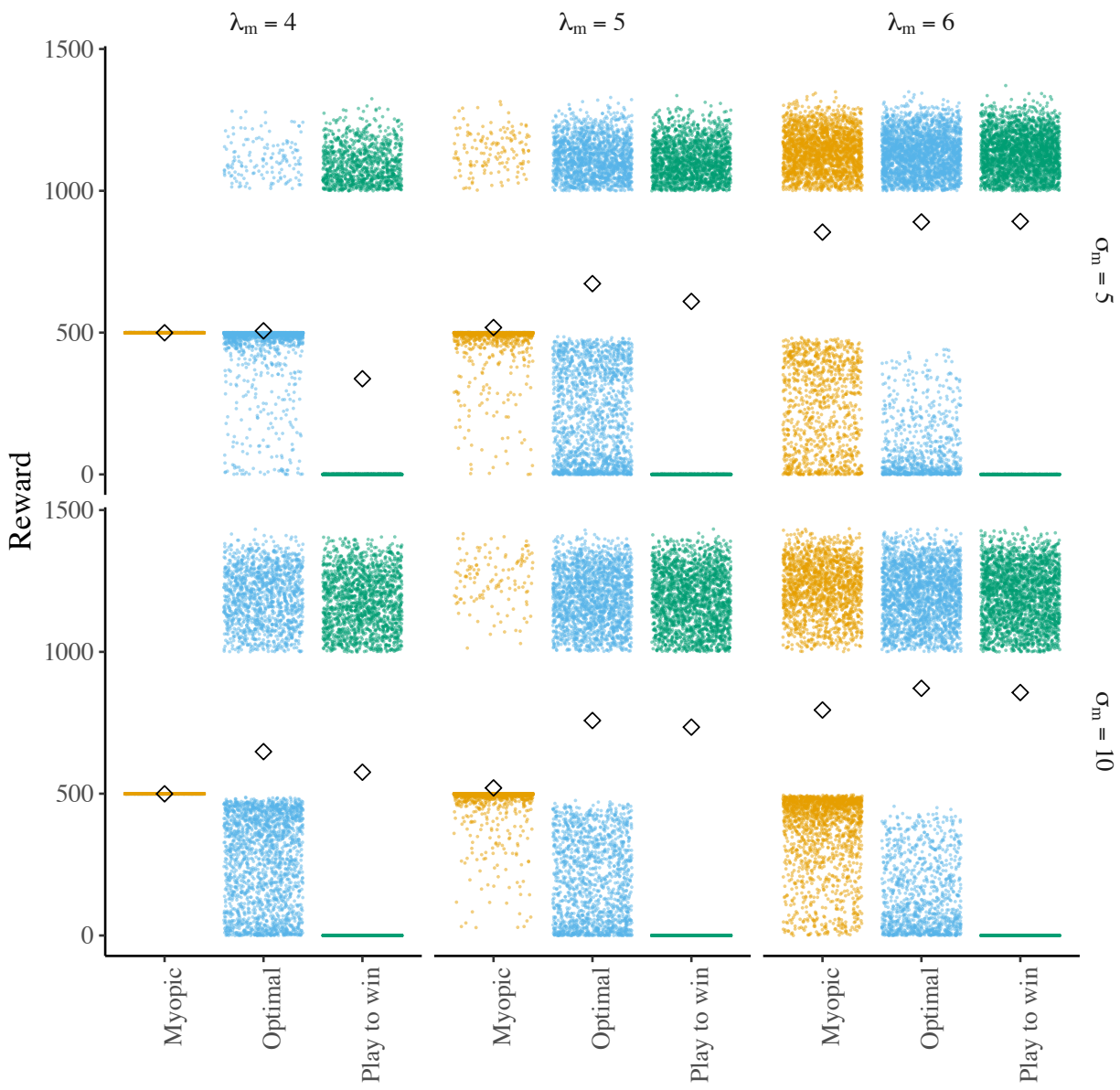

Figure 7: Visualization of the performance of the three strategies in terms of total reward earned as a function of performance variance σ_m and skill level λ_m, where the colored dots indicate the individual outcomes over 10,000 simulated instantiations of the Wiener process for each strategy. Other parameters are defaulted to $T = 10$, $\sigma_s = 5$, $B = 1000$, $\Delta = 50$, $\nu = 10$, and $\lambda_s = 5$. Diamonds indicate the average reward. Rewards above 1000 indicate that the agent succeeded in achieving the make-or-break goal and may also have earned an additional reward from the safe-reward task. Larger rewards indicate that the agent succeeded earlier and had more time to invest in the safe-reward task. Rewards equal to 500 indicate that the agent did not pursue the make-or-break goal at all or immediately reached the giving-up threshold, and thus allocated all available time to the safe task. Rewards equal to 0 indicate that the agent pursued the make-or-break task unsuccessfully, and exhausted all time resources without any reward. Rewards between 500 and 0 imply that the decision maker began investing in the make-or-break task, but gave up and switched to the safe-reward task.

it is only marginally profitable (from a myopic perspective) to allocate any amount of time to the make-or-break task. The myopic threshold is rather conservative, leading to quick dropouts in most simulation runs. Agents following the myopic strategy achieve success in very few simulation runs (3.1%). In contrast, simulated agents following the optimal strategy persevere for longer and succeed in approximately half the trials, with dropouts more evenly distributed across time (see the upper middle panel of Figure 7). At intermediate skill levels, the myopic strategy performs slightly better than playing it safe, whereas the optimal giving-up strategy achieves substantial rewards (average rewards of 651 for the optimal strategy vs. 512 for the myopic strategy). The difference between the two giving-up strategies is still notable—yet much smaller—at relatively high skill levels ($\sigma_m = 5, \lambda_m = 6$, see the upper right panel of Figure 7), and gradually declines as the value of λ_m increases, until the two strategies become indistinguishable (no dropouts are observed).

The effect of uncertainty: The two strategies diverge further in terms of observed behavior and expected returns as the level of uncertainty in the environment increases. First, higher uncertainty entails that the observed behavior for the two giving-up strategies deviates for a larger set of parameters, with investment in the make-or-break task being optimal for even lower skill levels. Second, higher uncertainty implies larger fluctuations in performance during the course of the task, leading myopic agents to prematurely drop out even at higher skill levels. As a result, the optimal strategy does much better than the myopic strategy in terms of gleaned rewards——at the same skill levels—than in environments with lower uncertainty (compare the upper and lower panels of Figure 7). For intermediate skill levels in the make-or-break task ($\lambda_m = 5$), rewards for the optimal giving-up strategy increase substantially in the high uncertainty environment (average reward of 726 for $\sigma = 10$ vs. 651 for $\sigma = 5$), whereas rewards for the myopic strategy are only marginally higher than playing it safe (average reward for the myopic strategy of 512 for both $\sigma = 5, 10$). The same holds for higher skill levels, where the myopic strategy leads to many early dropouts, whereas the distribution of dropouts for the optimal strategy is skewed toward later stages of the task (i.e., when most of the available time has already been allocated; see bottom right of Fig. 7, where $\lambda_m = 6$ and $\sigma_m = 10$). This is reflected in the expected returns on the strategies, where the average rewards for the myopic strategy are substantially lower than in lower uncertain environments (average reward of 800 for $\sigma = 10$ vs. 868 for $\sigma = 5$). In contrast, there was only a marginal decrease in the average rewards for the optimal strategy (average reward of 883 for $\sigma = 10$ vs. 896 for $\sigma = 5$).

When and why you should never give up. For moderate or high skill levels in the make-or-break task, the play-to-win heuristic does fairly well relative to the optimal strategy in terms of average rewards, and even outperforms the myopic strategy. As uncertainty in the environment increases, the play-to-win heuristic even approximates the expected rewards of the optimal strategy (see Figure 7, middle and right columns for $\sigma = 10$ and $\lambda_m = 5, 6$, where play-to-win achieved average rewards of 648 and 838 vs. 726 and 862 for the optimal strategy). How is this possible, given that the heuristic also eliminates computational effort? In environments with high uncertainty, the optimal strategy is itself more tolerant in order to accommodate the increased value of information. Thus, even in cases where performance falls below the optimal giving-up threshold, it is not unlikely that performance will rebound and ultimately reach the success threshold.

In environments with extreme uncertainty, the play-to-win heuristic approximates the expected rewards of the optimal giving-up strategy, rendering the computational costs for deriving the optimal threshold an unnecessary burden.

Although it is plausible that managers familiar with the stochastic processes literature will be able to apply the analytical formula to calculate the returns on the play-to-win strategy, it is unlikely that laypeople will do so. We therefore suggested a behaviorally plausible starting rule that executes a myopic calculation to assess whether it is worthwhile investing in the make-or-break task at all, in which case effort is continuously allocated until success or failure ensues. How does the play-to-win strategy perform when combined with such a starting rule? The starting rule would protect decision makers from investing in the make-or-break task in pathological cases, where the prospects of success seem impossibly bleak at the outset (see upper left panel of Figure 7 when $\sigma_m = 5$ and $\lambda_m = 4$, where the play-to-win strategy would succeed in only about 27% of the cases, and gain no rewards in other cases). The starting rule would also stop decision makers from investing in the make-or-break task in a small set of environments where it would actually be profitable (see bottom left panel of Figure 7). In sum, the play-to-win heuristic relying on a myopic starting rule outperforms the myopic giving-up strategy in all examined environments where the rule suggests that it is worth pursuing the make-or-break task, and especially so when uncertainty is high (see the center and right of Figure 7, contrast top and bottom). This result suggests that laypeople with limited computational capacities can leverage the smart potential of the play-to-win heuristic in most environments, and achieve high expected rewards by disregarding progress information (i.e., observed performance quality at any point in the task).

4 General Discussion

When should people invest time in risky but highly rewarding, make-or-break tasks, disregarding other activities with safe rewards? How much effort should a scientist invest towards applying for a prestigious grant? When should an entrepreneur give up on a high-risk yet potentially high-reward startup that is currently performing less well than expected? Researchers across disciplines in the social and behavioral sciences have alluded to the crucial factors involved in making these decisions (e.g., Eccles & Wigfield, 2002; Vroom, 1964), but the problem has previously evaded rigorous analysis. In this article, we present a formal model that addresses this conceptual gap and can be applied to effort or investment allocation problems in organizational, educational, and managerial settings.

4.1 On the large impact of small changes

We found that the expected returns of the make-or-break tasks are sigmoid in shape, which implies that similar problems encountered in real-life also harbor such non-convexities between investment and reward. Additionally, very small changes in the parameters of the problem can shift the prescribed allocation policy from one extreme (i.e., investing nothing) to the other (i.e., allocating all available resources; see Kukla, 1972; Vancouver, More, & Yoder, 2008). For instance, consider two students with similar skill levels studying in the same competitive program. The optimal strategy may prescribe one student to drop out entirely,

while the other should double down and invest all available resources into their studies. Similarly, relaxing the deadline for a grant application or reducing the threshold required to achieve a bonus at work may mobilize people to dedicate themselves fully to a task, whereas they might have appeared indifferent before. Crucially, we showed that the optimal allocation policy is moderated by the degree of uncertainty in the relation between investment and performance. In environments with high uncertainty, people with a larger range of skill levels will be motivated to pursue challenging goals, either in hope of an outside chance of success or working hard to minimize the offhand chances of failure. These results are at odds with the prevailing time allocation theories in economics, behavioral biology, and psychology, which typically assume that returns on different activities diminish as a function of the effort invested, or that different activities complement each other (e.g., Becker, 1965; Borjas & Van Ours, 2000; Charnov, 1976; Kurzban et al., 2013). In these theories, the underlying optimization problems are convex, with only a single optimum that shifts gradually as the parameters of the problem are varied. Likewise, the predicted policy switches cannot be accounted for by theories of motivation in cognitive and organizational psychology (Bandura & Locke, 2003; Heath et al., 1999; Locke & Latham, 2002), which suggest that people tend to respond to marginal changes in the difficulty of achieving a goal with continuous (either monotonic or non-monotonic) changes in the amount of effort exerted.

4.2 The computational challenges of dynamic decision making

There are two modes of discovery in the disciplines investigating human decision making. Mathematical psychologists and neuroscientists tend to start their inquiry by formulating behaviorally plausible cognitive models, and then conduct experiments to assess the extend to which people's behavior corresponds to different behavioral models. The descriptive, rather than the normative value of these models drives the inquiry. For example, they have extensively investigated dynamic decision-making in problems where people choose between two multi-attribute options, and where information about the value of the options is accessed and evaluated dynamically (e.g., by retrieving memories of similar items experienced in the past). This is also the case in our setting, where new information is assumed to unfold dynamically as a stochastic drift over time. Furthermore, these models assume that decision makers make a choice when the evidence accumulated in favor of an alternative has crossed a decision threshold (e.g., Busemeyer & Townsend, 1993; Khodadadi, Fakhari, & Busemeyer, 2017; Krajbich, Armel, & Rangel, 2010; Ratcliff et al., 2016; Usher & McClelland, 2001). One of the main advantages of these dynamic models is that they generate a rich set of predictions about the timing of events, namely, when and how often different decision thresholds will be crossed at different parameterizations of the problem. At the same time, it has been increasingly valuable to leverage dynamic optimization techniques to obtain insights in the properties of the optimal policies and to assess the conditions under which different action thresholds lead to good decisions (see Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Fudenberg, Strack, & Strzalecki, 2015; Malhotra et al., 2017; Tajima et al., 2016). Are the optimal policies for this task also cognitively plausible strategies? Our computational analysis suggests it is unlikely. Nevertheless, knowing the optimal policies can guide the search for computationally more cost-effective alternatives that can approximate optimality.

In contrast, economists and scientists in neighboring disciplines begin their inquiry from normative mod-

els. Researchers in economics and finance have extensively studied optimal policies in dynamic decision-making settings where the value of an asset—commonly described by a Wiener process—fluctuates over time (e.g., a stock, or the value of a factory) and agents are required to make high-impact and irreversible decisions (e.g., selling a factory; see McDonald & Siegel, 1986; Pindyck, 1991). These problems have clear-cut analogs in common problems encountered by laypeople, such as whether and when to sell a house or quit a stressful job. Dixit and Pindyck (1994) were well aware of the relevance of their modeling framework for everyday decisions and briefly discuss marriage as an irreversible dynamic decision that most people make at least (or at most) once. To date, only a handful of experimental studies have investigated how people make decisions in such dynamic settings, where the tasks are framed as investment decisions (Oprea, Friedman, & Anderson, 2009; Strack & Viefers, 2013). How do people cope with the computational complexity of dynamic decision making in these settings when the optimal strategy is computationally inaccessible? Defining and studying computationally simpler, boundedly rational strategies could lead to a rich set of predictions about behavior in dynamic decision environments (see, e.g., Hertwig, Davis, & Sulloway, 2002; Oprea et al., 2009).

The dynamic task we studied has structural and technical similarities with both these strands of research. Thinking like economists, we were able to characterize the optimal giving-up strategy for a family of common problems. In the spirit of Herbert Simon (1978), we employed complexity theory to gauge the computational requirements of the optimal strategy. We showed that it is computationally intractable for humans, and that it is computationally expensive even for modern computers. We then presented two psychologically plausible alternatives, motivated by the principles of myopic and heuristic decision making. These simpler strategies come with substantially cheaper computational costs. Deriving the optimal strategy enabled us to assess the prescriptive value of the boundedly rational strategies for different parameters of the decision-making problem. As in the dynamic decision-making problems investigated in mathematical psychology and neuroscience, our models generate a rich set of predictions about the timing of success and dropout at different parameterizations of the problem. Clearly, we have not yet exhausted the space of psychologically plausible stopping strategies (see Busemeyer & Rapoport, 1988). For instance, an agent may decide to stop when performance stagnates for a long period of time, constructing a local (in time) stopping rule. Formally analyzing additional strategies is beyond the scope of this paper, yet our framework paves the way for future experiments that will make it possible to assess the descriptive value of the proposed strategies and to identify other plausible psychological alternatives.

4.3 Risky choices, sunk costs, and bounded rationality

We started from the assumption that decision makers maximize expected value. This was a convenience assumption that allowed us to convey our results with the greatest clarity. How do our results connect to the literature on risky choice?[9] When the marginal utility of rewards diminishes (Bernoulli, 1954), we would expect people to discount the expected rewards from the two activities, especially from the large

[9]Note that some form of intertemporal discounting is often built into dynamic decision-making models. We avoided adding a discount factor to prevent the mathematics from appearing more daunting without adding much substance. Yet the complexity arguments we made about risk preferences also hold for time preferences (Berns, Laibson, & Loewenstein, 2007).

make-or-break bonus. Going a step further, people may subjectively reassess the value of succeeding or dropping out in the make-or-break task (Atkinson, 1957; Heath et al., 1999; Lewin et al., 1944) or distort the probabilities of success and failure at each level of invested effort (Kahneman & Tversky, 1979). How do these findings transfer to the complex and dynamic decision-making settings encountered in the real world? There is evidence that as the environment becomes more complex (e.g., the number of alternatives increases) people tend to rely more on computationally simple strategies (e.g., Venkatraman, Payne, & Huettel, 2014). Every variable transformation, such as those implied by expected and non-expected utility theories, has to be computed, either at a conscious or an unconscious level. However, we have already shown that the optimal strategy is computationally too expensive for humans, regardless of the mode of computation. Thus, deriving a risk-averse or time-discounted policy using backwards induction may have normative value (e.g., Smith & Ulu, 2017) but no descriptive value; such strategies would incur even larger computational costs than the original optimal strategy. Plausible theories of risky or intertemporal choice in dynamic contexts have to address the severe computational challenges of deriving consistent strategies. In two recent studies, Barberis (2012) and Ebert and Strack (2015) addressed this issue by assuming that some people are described by prospect theory, while at the same time taking risks myopically, as if they disregarded future choices.

In dynamic real-world problems, people may make risky choices primarily by responding to computational limitations that require them to find simple and robust strategies that trade off effectively between complexity and expected returns (Brandstätter, Gigerenzer, & Hertwig, 2006; J. W. Payne, Bettman, & Johnson, 1993). The boundedly rational strategies we have proposed produce different risk-taking patterns, highlighting another approach to the study of risky behavior in dynamic contexts. The myopic giving-up strategy, for instance, becomes more conservative as uncertainty in the environment increases, reproducing a behavioral pattern that could be derived by passing rewards through a risk-averse utility function and then calculating the optimal strategy, yet at a much lower computational cost. The play-to-win strategy, by contrast, produces behavior that is compatible with the sunk cost fallacy (Arkes & Blumer, 1985) and could also be seen as a version of risk-seeking. Thus, instead of subjectively transforming rewards and probabilities, people might express their risk tendencies by committing to boundedly rational strategies. Subjective transformations of value and distortions of perceived probabilities might be important components of dynamic choices, but they have to operate in tandem with another behavioral assumption—such as myopia—to be behaviorally plausible. We intend to investigate these issues in future experiments.

4.4 Self-efficacy and learning

When people embark on new, challenging projects (such as pursuing a PhD or founding a startup), they often only have a rough estimate of their own abilities. Bandura (1977) coined the term "self-efficacy" to describe the belief in one's ability to succeed, complete tasks, and reach goals. A mismatch between believed and actual skill levels can alter how people allocate their time to different tasks. Under-confidence can have the most dramatic effect, leading to self-fulfilling prophecies, where people who underestimate their abilities may never pursue a highly rewarding make-or-break goal (also see Hogarth & Karelaia, 2012). Without direct experience of successful achievement of a challenging goal, the decision maker may never rectify their initial beliefs. In the one-shot version of the task, overconfidence can lead decision makers to under-invest

or over-invest depending on their skill level. In the dynamic version of the task, however, overconfidence may encourage myopic individuals to persevere for longer, thereby counteracting the conservative effect of myopic giving up (Nozick, 1994).

How do people form beliefs about their own skill level and the relationship between effort, luck, and achievement in a task? They may use social comparisons or individual experiences in similar tasks to inform their beliefs about their skill levels (Bandura, 1977; Bol, de Vaan, & van de Rijt, 2018; Frank, 1935). Repeated interactions with a task or continuous feedback about performance may help them to hone their intuitions about how effort and luck jointly define the chances of achieving a goal (Weiner, 1985; Weiner & Kukla, 1970). Even if people are certain about their skill level, they may leverage repeated interactions with the same task to refine their strategies or choose among them (Erev & Barron, 2005; Rieskamp & Otto, 2006). Oprea et al. (2009), for instance, studied how people behave in optimal stopping problems, where they have to choose when to sell an asset. Although the optimal strategy is similar in terms of complexity with the strategy described here, it emerged that, after repeated interactions with the same task, people were able to approximate optimality.

4.5 The value of perseverance

A main insight from the optimal giving-up strategy we advanced is that optimally behaving agents should persevere before dropping out, and that the strategy should become more tolerant as the amount of uncertainty in the environment increases. This result echoes findings from the dynamic investing literature in economics, where agents should wait longer before making irreversible decisions, thus harnessing the value of dynamically unfolding information (Dixit & Pindyck, 1994). Duckworth and collaborators have shown that the ability to persevere and maintain effort in the face of failures is at least as important as intelligence for succeeding in one's field (Duckworth, Peterson, Matthews, & Kelly, 2007). The unexpected outcomes of real-world problems can often lead to discouragement and make people doubt their chances of success. Perseverance when pursuing a high-stakes, yet challenging goal might be even more valuable in real-world settings, precisely because of the additional uncertainties involved in estimating one's abilities. In such settings, simple strategies that persevere for longer than reasonable under the current set of beliefs (e.g., play-to-win as an extreme case of a strategy that never gives up) could in fact lead to better outcomes.

4.6 Conclusion

A common movie plot involves a protagonist pursuing a challenging but rewarding make-or-break goal, overcoming multiple frustrations and drawbacks to finally succeed. In real life, people may rationally abstain from investing in such make-or-break goals or give up after initial setbacks, thereby shifting their efforts to other activities. Knowing when to pursue a challenging goal and when to give up are hard decision-making problems, which are pervasive in everyday experience. Life can be complex at times, and simple strategies relying on myopic or heuristic principles might allow us to navigate the uncertainty of these problems, and perhaps, successfully achieve our goals.

5 Acknowledgments

We thank Daniel Barkoczi, Jerome Busemeyer, Michalis Drouvelis, Daniel Friedman, Chien-Ju Ho, Thorsten Joachims, Konstantinos Katsikopoulos, Emmanuel Kemel, Bobby Kleinberg, Amit Kothiyal, Thorsten Pachur, Rajiv Sethi, Lisa Son, Philipp Strack, Ryutaro Uchiyama and the participants of the Concepts and Categories seminar at NYU for their insightful comments. We are grateful to Susannah Goss for editing the manuscript. This research was supported in part through NSF Award IIS-1513692.

6 Code Availability Statement

The code for producing the results reported in the paper is available at https://osf.io/p2vur/.

References

Arkes, H. R., & Blumer, C. (1985). The psychology of sunk cost. *Organizational Behavior and Human Decision Processes*, *35*(1), 124–140.

Atkinson, J. W. (1957). Motivational determinants of risk-taking behavior. *Psychological Review*, *64*(6), 359–372.

Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191–215.

Bandura, A., & Locke, E. A. (2003). Negative self-efficacy and goal effects revisited. *Journal of Applied Psychology*, *88*(1), 87–99.

Barberis, N. (2012). A model of casino gambling. *Management Science*, *58*(1), 35–51.

Bass, R. F. (2011). *Stochastic processes* (Vol. 33). Cambridge, United Kingdom: Cambridge University Press.

Becker, G. S. (1965). A theory of the allocation of time. *The Economic Journal*, *75*(299), 493–517.

Bellman, R. (2013). *Dynamic programming*. Princeton, NJ: Courier Corporation.

Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. *Econometrica*, *22*(1), 23–36.

Berns, G. S., Laibson, D., & Loewenstein, G. (2007). Intertemporal choice: toward an integrative framework. *Trends in Cognitive Sciences*, *11*(11), 482–488.

Bertsekas, D. P. (1995). *Dynamic programming and optimal control*. Belmont, MA: Athena Scientific.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700–765.

Bol, T., de Vaan, M., & van de Rijt, A. (2018). The matthew effect in science funding. *Proceedings of the National Academy of Sciences*. Retrieved from `http://www.pnas.org/content/early/2018/04/18/1719557115` doi: 10.1073/pnas.1719557115

Borjas, G. J., & Van Ours, J. C. (2000). *Labor economics* (Vol. 2). Boston, MA: McGraw-Hill.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge, United Kingdom: Cambridge University Press.

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: making choices without trade-offs. *Psychological Review*, *113*(2), 409–432.

Brent, R. P. (1976). The complexity of multiple-precision arithmetic. In R. S. Anderssen & R. P. Brent (Eds.), *The complexity of computational problem solving* (pp. 126–165). Brisbane, Australia: University of Queensland Press.

Busemeyer, J. R., & Rapoport, A. (1988). Psychological models of deferred decision making. *Journal of Mathematical Psychology*, *32*(2), 91–134.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*(3), 432–459.

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology*, *9*(2), 129–136.

Chhikara, R. S. (1989). *The inverse gaussian distribution: theory, methodology, and applications*. New York, NY: M. Dekker.

DeGroot, M. H. (n.d.). *Optimal statistical decisions*. Hoboken, NJ: John Wiley & Sons.

Diecidue, E., & Van De Ven, J. (2008). Aspiration level, probability of success and failure, and expected utility. *International Economic Review*, *49*(2), 683–700.

Dixit, A. K., & Pindyck, R. S. (1994). *Investment under uncertainty*. Princeton, NJ: Princeton University Press.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, *92*(6), 1087–1101.

Ebert, S., & Strack, P. (2015). Until the bitter end: on prospect theory in a dynamic context. *The American Economic Review*, *105*(4), 1618–1633.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*(1), 109–132.

Edwards, W. (1953). Probability-preferences in gambling. *The American Journal of Psychology*, *66*(3), 349–364.

Edwards, W. (1954). Probability-preferences among bets with differing expected values. *The American Journal of Psychology*, *67*(1), 56–67.

Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological review*, *112*(4), 912–931.

Feather, N. T. (1959). Subjective probability and decision under uncertainty. *Psychological Review*, *66*(3), 150–164.

Fechner, H. B., Schooler, L. J., & Pachur, T. (2018). Cognitive costs of decision-making strategies: A resource demand decomposition analysis with a cognitive architecture. *Cognition*, *170*, 102–122.

Festinger, L. (1942). A theoretical interpretation of shifts in level of aspiration. *Psychological Review*, *49*(3), 235–250.

Frank, J. D. (1935). The influence of the level of performance in one task on the level of aspiration in another. *Journal of Experimental Psychology*, *18*(2), 159–171.

Fudenberg, D., Strack, P., & Strzalecki, T. (2015). Stochastic choice and optimal sequential sampling.

Available at arXiv:1505.03342v1.

Gabaix, X., Laibson, D., Moloche, G., & Weinberg, S. (2006). Costly information acquisition: experimental analysis of a boundedly rational model. *The American Economic Review*, *96*(4), 1043–1068.

Gaissmaier, W., & Schooler, L. J. (2008). The smart potential behind probability matching. *Cognition*, *109*(3), 416–422.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451–482.

Heath, C., Larrick, R. P., & Wu, G. (1999). Goals as reference points. *Cognitive Psychology*, *109*(1), 79–109.

Heider, F. (1958). *The psychology of interpersonal relations*. Wiley.

Hertwig, R., Davis, J. N., & Sulloway, F. J. (2002). Parental investment: how an equity motive can produce inequality. *Psychological Bulletin*, *128*(5), 728–745.

Hey, J. D. (1982). Search for rules for search. *Journal of Economic Behavior & Organization*, *3*(1), 65–81.

Hogarth, R. M., & Karelaia, N. (2012). Entrepreneurial success and failure: Confidence and fallible judgment. *Organization Science*, *23*(6), 1733–1747.

Hoppe, F. (1931). Untersuchungen zur handlungs- und affektpsychologie. *Psychologische Forschung*, *14*(1), 1–62.

Hotaling, J. M., & Busemeyer, J. R. (2012). Dft-d: a cognitive-dynamical model of dynamic decision making. *Synthese*, *189*(1), 67–80.

Huys, Q. J., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S. J., ... Roiser, J. P. (2015). Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences*, *112*(10), 3098–3103.

Jacka, S. (1991). Optimal stopping and the American put. *Mathematical Finance*, *1*(2), 1–14.

Johnson, E., Camerer, C., Sen, S., & Rymon, T. (2002). Detecting failures of backward induction: monitoring information search in sequential bargaining. *Journal of Economic Theory*, *104*(1), 16–47.

Johnson, E., & Payne, J. W. (1985). Effort and accuracy in choice. *Management Science*, *31*(4), 395–414.

Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, *47*(2), 263–291.

Karatzas, I., & Shreve, S. (2012). *Brownian motion and stochastic calculus*. New York, NY: Springer Science & Business Media.

Khodadadi, A., Fakhari, P., & Busemeyer, J. R. (2017). Learning to allocate limited time to decisions with different expected outcomes. *Cognitive Psychology*, *95*, 17–49.

Kool, W., & Botvinick, M. (2014). A labor/leisure tradeoff in cognitive control. *Journal of Experimental Psychology: General*, *143*(1), 131–141.

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*(10), 1292–1298.

Kukla, A. (1972). Foundations of an attributional theory of performance. *Psychological Review*, *79*(6), 454–470.

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective

effort and task performance. *Behavioral and Brain Sciences*, *36*(06), 661–679.

Kushner, H., & Dupuis, P. G. (2013). *Numerical methods for stochastic control problems in continuous time*. New York, NY: Springer Science & Business Media.

Lancaster, T. (1972). A stochastic model for the duration of a strike. *Journal of the Royal Statistical Society. Series A (General)*, *135*(2), 257–271.

Lee, M.-L. T., & Whitmore, G. A. (2006). Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. *Statistical Science*, *21*(4), 501–513.

Lewin, K. (1936). Pyschology of success and failure. *Occupations: The Vocational Guidance Journal*, *14*(9), 926–930.

Lewin, K., Dembo, T., Festinger, L., & Sears, P. S. (1944). Level of aspiration. In J. M. Hunt (Ed.), *Personality and behavior disorders.* New York, NY: Ronald Press.

Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review*, *124*(6), 762–794.

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *The American Psychologist*, *57*(9), 705–717.

Malhotra, G., Leslie, D. S., Ludwig, C. J., & Bogacz, R. (2017). Time-varying decision boundaries: insights from optimality analysis. *Psychonomic Bulletin & Review*. Advance online publication.

McCardle, K. (1985). Information acquisition and the adoption of new technology. *Management Science*, *31*(11), 1372–1389.

McCardle, K., Tsetlin, I., & Winkler, R. (2016). When to abandon a research project and search for a new one. INSEAD Working Paper No. 2016/10/DSC. Available at SSRN: https://ssrn.com/abstract=2739699.

McDonald, R., & Siegel, D. (1986). The value of waiting to invest. *The Quarterly Journal of Economics*, *101*(4), 707–727.

McKelvey, R. D., & Palfrey, T. R. (1992). An experimental study of the centipede game. *Econometrica*, *60*(4), 803–836.

Meehl, P. E., & MacCorquodale, K. (1951). Some methodological comments concerning expectancy theory. *Psychological Review*, *58*(3), 230233.

Morvan, C., & Maloney, L. T. (2012). Human visual search does not maximize the post-saccadic probability of identifying targets. *PLoS Computational Biology*, *8*(2), e1002342.

Nozick, R. (1994). *The nature of rationality*. Princeton, NJ: Princeton University Press.

Oprea, R., Friedman, D., & Anderson, S. T. (2009). Learning to wait: a laboratory investigation. *The Review of Economic Studies*, *76*(3), 1103–1124.

Papadimitriou, C. H. (2003). *Computational complexity*. Chichester, United Kingdom: John Wiley & Sons.

Payne, J. W., Bettman, J., & Johnson, E. (1993). *The adaptive decision maker*. Cambridge, United Kingdom: Cambridge University Press.

Payne, J. W., Laughhunn, D. J., & Crum, R. (1980). Translation of gambles and aspiration level effects in risky choice behavior. *Management Science*, *26*(10), 1039–1060.

Payne, S. J., Duggan, G. B., & Neth, H. (2007). Discretionary task interleaving: heuristics for time allocation

in cognitive foraging. *Journal of Experimental Psychology: General*, *136*(3), 370–388.

Peskir, G., & Shiryaev, A. (2006). *Optimal stopping and free-boundary problems*. Basle, Switzerland: Birkhäuser.

Pindyck, R. S. (1991). Irreversibility, uncertainty, and investment. *Journal of Economic Literature*, *29*(3), 1110–1148.

Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review*, *106*(4), 643–675.

Pratt, J. (1964). Risk aversion in the small and in the large. *Econometrica*, *32*(1/2), 122–136.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281.

Redner, S. (2001). *A guide to first-passage processes*. Cambridge University Press.

Rieskamp, J., & Otto, P. E. (2006). Ssl: a theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*(2), 207–236.

Rotter, J. B. (1942). Level of aspiration as a method of studying personality. a critical review of methodology. *Psychological Review*, *49*(5), 463–474.

Shiryaev, A. N. (2007). *Optimal stopping rules*. New York, NY: Springer Science & Business Media.

Siegel, S. (1957). Level of aspiration and decision making. *Psychological Review*, *64*(4), 253–262.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*(1), 99–118.

Simon, H. A. (1978). Rationality as process and as product of thought. *The American Economic Review*, *68*(2), 1–16.

Smith, J. E., & Ulu, C. (2017). Risk aversion, information acquisition, and technology adoption. *Operations Research*, *65*(4), 1011–1028.

Son, L. K., & Sethi, R. (2006). Metacognitive control and optimal learning. *Cognitive Science*, *30*(4), 759–774.

Stojic, H., Analytis, P. P., & Speekenbrink, M. (2015). Human behavior in contextual multi-armed bandit problems. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 2290–2295). Austin, TX: Cognitive Science Society.

Strack, P., & Viefers, P. (2013). Too proud to stop: Regret in dynamic decisions. Available at SSRN: https://ssrn.com/abstract=2465840.

Tajima, S., Drugowitsch, J., & Pouget, A. (2016). Optimal policy for value-based decision-making. *Nature Communications*, *7*.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, *185*(4157), 1124–1131.

Ulu, C., & Smith, J. E. (2009). Uncertainty, information acquisition, and technology adoption. *Operations Research*, *57*(3), 740–752.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological review*, *108*(3), 550–592.

Vancouver, J. B., More, K. M., & Yoder, R. J. (2008). Self-efficacy and resource allocation: support for a nonmonotonic, discontinuous model. *Journal of Applied Psychology*, *93*(1), 35–47.

Van Rooij, I. (2008). The tractable cognition thesis. *Cognitive Science*, *32*(6), 939–984.

Varian, H. R. (1978). *Microeconomic analysis*. New York, NY: WW Norton.

Venkatraman, V., Payne, J. W., & Huettel, S. A. (2014). An overall probability of winning heuristic for complex risky decisions: choice and eye fixation evidence. *Organizational Behavior and Human Decision Processes*, *125*(2), 73–87.

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

Vroom, V. H. (1964). *Work and motivation*. New York, NY: John Wiley & Sons.

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, *16*(2), 117–186.

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review*, *92*(4), 548.

Weiner, B., & Kukla, A. (1970). An attributional analysis of achievement motivation. *Journal of Personality and Social Psychology*, *15*(1), 1.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Exploration and generalization in vast spaces. *bioRxiv*. doi: 10.1101/171371

Zhang, S., & Yu, J. A. (2013). Forgetful Bayes and myopic planning: human learning and decision-making in a bandit setting. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 26 (NIPS 2013)* (pp. 2607–2615). Red Hook, NY: Curran Associates.

7 Supplementary Material

7.1 The expected rewards of the make-or-break task is a sigmoid function of time

Recall that the expected reward from investing time t_m in the make-or-break task is

$$\mathbb{E}[r_m(t_m)] = B \cdot \left(1 - \Phi\left(\frac{\Delta - \lambda_m t_m}{\sigma_m \sqrt{t_m}}\right)\right). \tag{17}$$

We are now going to show that this is a sigmoid function for $t_m > 0$, meaning that $\mathbb{E}[r_m(t_m)]$ converges to a number $c \in \mathbb{R}$ as $t_m \to \infty$ and that it has a unique inflection point, with its second derivative being positive on the left and negative on the right of this point. The first statement is straightforward, because as $t_m \to \infty$, $\frac{\Delta - \lambda_m t_m}{\sigma_m \sqrt{t_m}} \to -\infty$, so $\mathbb{E}(r_m(t_m)) \to B$.

For the second statement, we write

$$\mathbb{E}[r_m(t_m)] = B \cdot \left(1 - \Phi\left(\frac{\Delta}{\sigma_m} t_m^{-\frac{1}{2}} - \frac{\lambda_m}{\sigma_m} t_m^{\frac{1}{2}}\right)\right) = B \cdot \left(1 - \Phi\left(a t_m^{-\frac{1}{2}} - b t_m^{\frac{1}{2}}\right)\right), \tag{18}$$

where $a = \frac{\Delta}{\lambda_m}$ and $b = \frac{\lambda_m}{\sigma_m}$. Therefore, it is enough to show that the function

$$f(t) = \Phi\left(a t^{-\frac{1}{2}} - b t^{\frac{1}{2}}\right), \;\; a, b > 0 \tag{19}$$

has a unique inflection point for $t > 0$, with its second derivative passing from a negative to a positive value. We have

$$f'(t) = -\frac{1}{2\sqrt{2\pi}} e^{-\frac{1}{2} \cdot \left(a t^{-\frac{1}{2}} - b t^{\frac{1}{2}}\right)^2} \cdot \left(a t^{-\frac{3}{2}} + b t^{-\frac{1}{2}}\right) \tag{20}$$

and

$$f''(t) = \frac{1}{4\sqrt{2\pi} \cdot t^{\frac{7}{2}}} \cdot e^{-\frac{1}{2} \cdot \left(a t^{-\frac{1}{2}} - b t^{\frac{1}{2}}\right)^2} \cdot \left(b^3 t^3 + \left(ab^2 + b\right) t^2 + \left(3a - a^2 b\right) t - a^3\right) \tag{21}$$

Because the first two factors in the expression for $f''(t)$ are always positive, its sign is determined by the third degree polynomial

$$z(t) = b^3 t^3 + \left(ab^2 + b\right) t^2 + \left(3a - a^2 b\right) t - a^3. \tag{22}$$

It is therefore enough to show that $z(t)$ has exactly one root for $t > 0$ and its sign switches from negative to positive as t crosses that root from left to right.

By further noticing that $z(0) < 0$ and $z(t) \to \infty$ as $t \to \infty$, it is sufficient to show that $z(t)$ has at most one local extremum for $t > 0$ (which will have to be a local minimum), or that $z'(t)$ has at most one root for $t > 0$. We calculate

$$z'(t) = 3b^3 t^2 + 2(ab^2 + b)t + 3a - a^2 b. \tag{23}$$

Even if this equation has two roots, their sum has to be equal to $-\frac{2(ab^2+b)}{3b^3}$, so that at least one has to be negative. This concludes the proof that $\mathbb{E}[r_m(t_m)]$ is a sigmoid function.

7.2 Optimal time allocation policy for the one-shot problem

In this section we study how the optimal policy for the one-shot allocation task varies with some of the parameters of the problem. We begin by studying some general properties of the total expected reward function, which is given by

$$h(t_m, t_s) = \mathbb{E}[r_m(t_m)] + \mathbb{E}[r_s(t_s)], \tag{24}$$

where t_m is the time invested in the make-or-break task and t_s is the time invested in the safe-reward task. If the total available time is T, then we may write $t_s = T - t_m$, hence the above simplifies to

$$h(t_m) = \mathbb{E}[r_m(t_m)] + \mathbb{E}[r_s(T - t_m)]. \tag{25}$$

The optimal amount of time invested in the make-or-break task is then given by

$$t_{opt} = \underset{0 \le t_m \le T}{\arg\max}\{h(t_m)\} \tag{26}$$

and the expected reward for this optimal strategy is $h(t_{opt})$. In case the maximum is attained at more than one point, we interpret arg max to be the first of these points.[10]

Substituting the equations for the expected rewards into Equation 25, we obtain

$$h(t_m) = B \cdot \left(1 - \Phi\left(\frac{\Delta - \lambda_m t_m}{\sigma_m \sqrt{t_m}}\right)\right) + v \cdot \lambda_s \cdot (T - t_m), \tag{27}$$

for $t_m \in (0, T]$ and $h(0) = v \cdot \lambda_s \cdot T$. For the first derivative we have

$$h'(t_m) = \frac{B}{2\sqrt{2\pi} \cdot \sigma_m} \cdot \left(\lambda_m \cdot t_m^{-\frac{1}{2}} + \Delta \cdot t_m^{-\frac{3}{2}}\right) \cdot e^{-\frac{(\Delta - \lambda_m t_m)^2}{2\sigma_m^2 t_m}} - v \cdot \lambda_s, \tag{28}$$

for $t_m \in (0, T]$ and $h'(0) = -v \cdot \lambda_s$. For the second derivative, we have $h''(t_m) = [\mathbb{E}(r_m(t_m))]'$, which we showed in the previous section has a single root for $t_m > 0$.

Although we are only interested in times $t_m \le T$, in studying how t_{opt} varies with the parameters T, λ_m, and B, it will be useful to consider the function $h(t_m)$ for any $t_m \ge 0$. We will need the following lemmas:

Lemma 7.1. *The function $h(t_m)$ either has no local extrema on $(0, \infty)$, or it has a unique local minimum at t_{min} and a unique local maximum at t_{max}, with $t_{min} < t_{max}$. In the latter case, its derivative $h'(t_m)$ is strictly positive inside the interval (t_{min}, t_{max}) and strictly negative outside.*

Proof: Note that $h'(0) < 0$. Therefore, if $h(t_m)$ must have a local minimum before any local maximum in $(0, \infty)$. Also, recall that $h''(t_m)$ has at most one root. Therefore, by the Mean Value Theorem (MVT), $h'(t_m)$ cannot have more than two roots. We conclude that $h(t_m)$, if it has any local extrema, has exactly one local minimum at t_{min} and one local maximum at t_{max}, with $t_{min} < t_{max}$, and that $h'(t_m)$ is positive on the interval (t_{min}, t_{max}) and negative on $(0, t_{min})$ and (t_{max}, ∞). □

[10] Continuity of $h(t_m)$ guarantees that a "first" such point exists.

Lemma 7.2. *When the local extrema of $h(t_m)$ exist, their positions t_{min} and t_{max} vary continuously with the parameters B, σ_m, λ_m, λ_s, v, Δ, and are independent of T.*

Proof: Independence from T follows from the fact that $h'(t_m)$ is independent of T. For the other statement, note that by the MVT, the unique root of $h''(t_m)$ must occur in the interval (t_{min}, t_{max}). This implies that $h''(t_{min}) > 0$ and $h''(t_{max}) < 0$. Therefore, by continuity of $h''(t_m)$ and of the (partial) derivatives of h' with respect to any of the parameters, we conclude that both local extrema positions t_{min} and t_{max} are stable, in the sense that they persist for small changes of the parameters and vary continuously with them. □

Lemma 7.3. *i) If $h(t_m) \geq h(0)$ for some $t_m \in (0,\infty)$, then $h(t_m)$ has a unique local maximum at $t_{max} \in (0,\infty)$, which is also global.*

ii) If $t_{opt} \neq 0$, then the condition in (i) holds and, moreover, $t_{opt} = \min\{T, t_{max}\}$.

Proof:

i) Since $h'(t_m) < 0$ for large t_m, $h(t_m)$ attains a global maximum in $[0,\infty)$. Even if that happens to be at 0, then by assumption it must also be attained somewhere in $(0,\infty)$. This global maximum in $(0,\infty)$ will also be a local maximum, which by Lemma 7.1 is unique.

ii) If $t_{opt} \neq 0$, then by definition there exists some $t_m \in (0,T]$, such that $h(t_m) > h(0)$. From part (i), there exists a unique local maximum $t_{max} \in (0,\infty)$. Clearly, if $t_{max} \leq T$, then $t_{opt} = t_{max}$. If on the other hand $t_{max} > T$, then since $h(t_m)$'s unique local maximum is at t_{max}, it has no local maximum in $(0,T)$, so its maximum in the interval $[0,T]$ is attained either at 0 or at T, which means that $t_{opt} = 0$ or $t_{opt} = T$. By assumption, the first possibility is excluded, therefore $t_{opt} = T$. This concludes the proof that $t_{opt} = \min\{T, t_{max}\}$. □

7.2.1 Varying the total available time T

Suppose first that we vary the total available time T, while all other parameters are kept fixed. We want to look at how t_{opt} varies. We distinguish two cases.

First suppose that $h(0) \geq h(t_m)$ for all $t_m \in [0,\infty)$. Note from Equation 27 that this condition is independent of T. If it is true, then investing zero time in the make-or-break task is always preferable to investing non-zero time in it. In other words, $t_{opt} = 0$ no matter what the value of T, so this case is trivial.

For the other case, we have the following proposition.

Proposition 7.4. *Suppose that $h(t_m) > h(0)$ for some $t_m > 0$. Then,*

$$t_{opt} = \begin{cases} 0, & \text{if } T \leq T^* \\ T, & \text{if } T^* < T \leq t_{max} \\ t_{max}, & \text{if } T > t_{max} \end{cases}, \tag{29}$$

where t_{max} is the unique global maximum of $h(t_m)$ and T^* is given by

$$T^* = \inf\{t_m \in (0,\infty) : h(t_m) > h(0)\}, \tag{30}$$

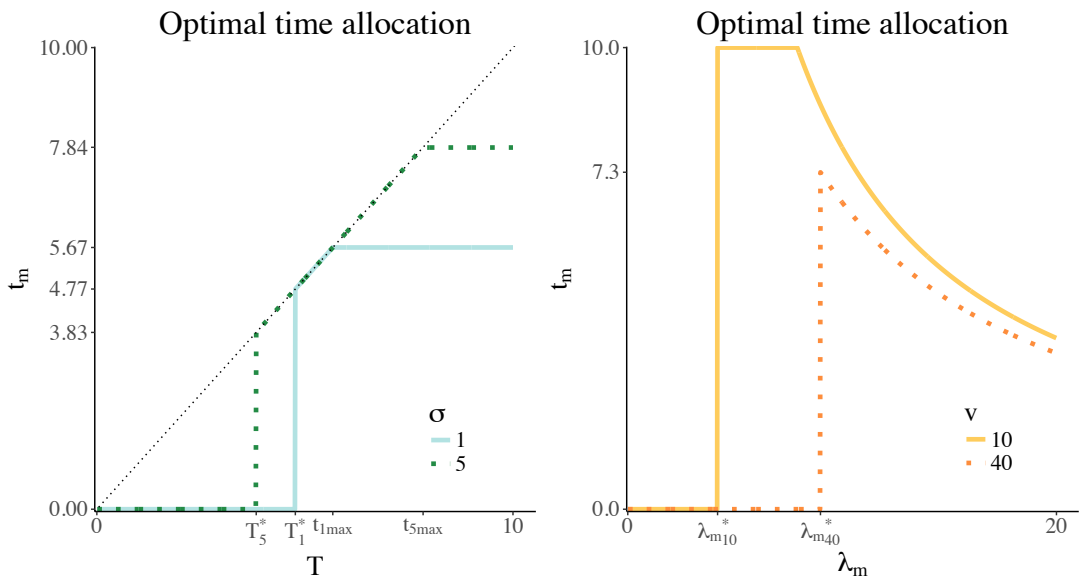

Figure 8: **Left:** The optimal amount of time allocated to the make-or-break task as a function of the total available time $T \in [0,10]$. At very low amounts of total available time (i.e., a tight deadline), the agent allocates all their time to the safe alternative. At T^*, the agent puts all their effort into the make-or-break task. As the total available time increases, the agent continues to invest all their time in the make-or-break task to increase the chances of achieving it, until the point t_{max}, at which the benefits from improving the chances of success are offset by the opportunity cost. Beyond that point, the agent should invest all available time in the safe task, which entails that the time allocation problem has an internal solution. We illustrate these time allocation patterns for low and intermediate uncertainty ($\sigma = 1$ vs. $\sigma = 5$) and denote time allocated to the make-or-break task by a subscript. **Right:** The optimal amount of time allocated to the make-or-break task as a function of the skill level in the task $\lambda_m \in [0,20]$. At skill level λ_m^*, the agent should discontinuously change their strategy from allocating all their time to the safe-reward task to allocating all or at least some of their time to the make-or-break task. When the opportunity cost of time is relatively low ($v = 10$), the agent should invest all their time in the make-or-break task; when the opportunity cost ($v = 40$) is relatively high, some of their time. As the skill level continues to increase, the proportion of time the agent should optimally allocate to the make-or-break task decreases. Consistently with the main text, in both panels we use the default parameter values of $T = 10$, $\sigma_m = \sigma_s = 5$, $B = 1000$, $\Delta = 50$, $v = 10$, $\lambda_m = 10$. We default λ_s to 3 for illustrative purposes and to ensure consistency with Figure 4 in the main text.

and satisfies $T^* < t_{max}$.

Proof: Let T^* be as in Equation 30. By continuity, $h(T^*) = h(0)$, so $t_{opt} = 0$ if and only if $T \leq T^*$. For $T > T^*$, by Lemma 7.1 we have $t_{opt} = \min\{T, t_{max}\}$. Finally, note that since $h(t_{max}) > h(0)$, we have $T^* < t_{max}$. □

7.2.2 Varying the skill level λ_m

We now study how t_{opt} changes as we vary λ_m, assuming that all other parameters remain constant. In this section, we use the notation $h(t_m; \lambda_m)$ and $t_{opt}(\lambda_m)$ in order to emphasize the dependence of these quantities on λ_m. For $t_m = 0$, $h(0; \lambda_m)$ does not depend on λ_m, so we will write just $h(0)$.

To simplify the presentation of the result, we will assume that when the skill level for the make-or-break task is 0, then the optimal strategy is always to invest all available time into the safe-reward task.[11]

[11] This does not have to be the case, especially if the reward threshold Δ is low. If it is not true, then we can show that $t_{opt}(\lambda_m)$ is always positive and varies continuously with λ_m.

Mathematically this means that $h(t_m;0) < h(0)$ for all $t_m > 0$.

We show that when λ_m is smaller than some value λ_m^*, then the optimal policy is to allocate all time to the safe-reward task, that is $t_{opt}(\lambda_m) = 0$. For $\lambda_m > \lambda_m^*$, the optimal policy is to put at least some time into the make-or-break task. Moreover, the transition at λ_m^* is discontinuous, with t_{opt} jumping from 0 to $t^* > 0$. As λ_m increases further, $t_{opt}(\lambda_m)$ will change continuously (but never increase).

The value of t^* can be either T, in which case at λ_m^* there will be a transition from allocating all time to the safe-reward task to allocating all time to the make-or-break task, or it may be smaller than T, in which case the transition will be from allocating all time to the safe-reward task to allocating time to both tasks. In the following proposition, we prove all of the above and give some technical conditions that tell us whether $t^* = T$ or $t^* < T$.

Proposition 7.5. *Suppose that $h(t_m;0) < h(0)$ for all $t_m \in (0,T]$. We have the following:*

i)

$$t_{opt}(\lambda_m) = \begin{cases} 0, & \text{if } \lambda_m \leq \lambda_m^* \\ \min\{T, t_{max}(\lambda_m)\}, & \text{if } \lambda_m > \lambda_m^* \end{cases}, \tag{31}$$

where

$$\lambda_m^* = \min\{\lambda_m > 0 : \exists t_m \in (0,T], h(t_m;\lambda_m) = h(0)\}. \tag{32}$$

ii) For $\lambda_m > \lambda_m^*$, $t_{opt}(\lambda_m)$ is a continuous function of λ_m, and

$$\lim_{\lambda_m \searrow \lambda_m^*} t_{opt}(\lambda_m) = t^* \tag{33}$$

where $t^* = \min\{T, t_{max}(\lambda_m^*)\}$. Moreover, $t^* = T$ if and only if $h'(T;\lambda_m^*) \geq 0$.

Proof:

i) For any $t_m \in (0,T]$, we have that $h(t_m;0) < h(0)$, $h(t_m;\lambda)$ is strictly increasing in λ_m and $\lim_{\lambda_m \to \infty} h(t_m;\lambda_m) > h(0)$. Therefore, the equation $h(t_m;\lambda_m) = h(0)$ has a unique solution. We define $\overline{\lambda}_m(t_m)$ to be this solution and note that $h(t_m;\lambda_m) > h(0)$ if and only if $\lambda_m > \overline{\lambda}_m(t_m)$.

Because the first partial derivatives of $h(t_m;\lambda_m)$ with respect to t_m and λ_m are both continuous and the one with respect to λ_m is non-zero, the Implicit Function Theorem implies that the solution $\overline{\lambda}_m(t_m)$ of $h(t_m;\lambda_m) = h(0)$ is a continuously differentiable function of t_m. Therefore, $\overline{\lambda}_m(t_m)$ attains a minimum in every interval of the form $[c,T]$, with $c > 0$. We will show that it also attains a minimum in $(0,T]$.

Note that for any given $\lambda_m > 0$, $h'(0;\lambda_m) < 0$. Therefore, by continuity of $h'(t_m;\lambda_m)$, there exists some $\varepsilon = \varepsilon(\lambda_m) > 0$, such that $h(t_m;\lambda_m) < h(0;\lambda_m) = h(0)$ for any $t_m \in (0,\varepsilon)$. In particular, there exists some $\varepsilon > 0$, such that $h(t_m;\overline{\lambda}_m(T)) < h(0)$, hence also $\overline{\lambda}_m(t_m) > \overline{\lambda}_m(T)$, for any $t_m < \varepsilon$. This proves our claim that $\overline{\lambda}_m(t_m)$ attains a minimum in $(0,T]$.

We define

$$\lambda_m^* = \min_{t_m \in (0,T]} \{\overline{\lambda}_m(t_m)\} = \min\{\lambda_m > 0 : \exists t_m \in (0,T], h(t_m;\lambda_m) = h(0)\}. \tag{34}$$

Recalling that $h(0) \geq h(t_m;\lambda_m)$ if and only if $\lambda_m \leq \overline{\lambda}_m(t_m)$, we obtain

$$t_{opt}(\lambda_m) = 0 \Leftrightarrow h(0) \geq \sup_{0<t_m\leq T} h(t_m;\lambda_m) \Leftrightarrow \lambda_m \leq \min_{0<t_m\leq T} \overline{\lambda}_m(t_m) = \lambda_m^*. \tag{35}$$

Combining this with Lemma 7.3, the result follows.

ii) Continuity of $t_{opt}(\lambda_m)$ for $\lambda_m > \lambda_m^*$ follows from part (i) and Lemma 7.2.

From the definition of λ_m^*, it follows that there exists some $t_m \in (0,T]$ such that $h(t_m;\lambda_m^*) = h(0)$. By Lemma 7.3, $h(t_m;\lambda_m)$ has a unique local maximum $t_{max}(\lambda_m^*)$, and by Lemma 7.2,

$$\lim_{\lambda_m \to \lambda_m^*} t_{max}(\lambda_m) = t_{max}(\lambda_m^*). \tag{36}$$

Combining this with part (i), we obtain

$$\lim_{\lambda_m \searrow \lambda_m^*} t_{opt}(\lambda_m) = \min\left\{\lim_{\lambda_m \searrow \lambda_m^*} t_{max}(\lambda_m), T\right\} = \min\{t_{max}(\lambda_m^*), T\} > 0, \tag{37}$$

because both T and $t_{max}(\lambda_m^*)$ are greater than zero.

Finally, note that by Lemma 7.1 we have that $h'(T;\lambda_m^*) \geq 0$ if and only if $t_{min}(\lambda_m^*) \leq T \leq t_{max}(\lambda_m^*)$. But $T < t_{min}(\lambda_m^*)$ is impossible anyway, because then $h(t_m;\lambda_m^*)$ would be strictly decreasing in $[0,T]$, contradicting the definition of λ_m^*. Therefore, $h'(T;\lambda_m^*) \geq 0$ if and only if $T \leq t_{max}(\lambda_m^*)$. Combining this with Equation 37, we get that $h'(T;\lambda_m^*) \geq 0$ if and only if $\lim_{\lambda_m \searrow \lambda_m^*} t_{opt}(\lambda_m) = T$.

□

7.2.3 Varying the reward B

We now study how t_{opt} changes as we vary B, assuming that all other parameters remain constant. In this section, we use the notation $h(t_m;B)$ and $t_{opt}(B)$ in order to emphasize the dependence of these quantities on B. The results and the proof are very similar as for λ_m, with one exception: here, the condition $h(t_m;0) < h(0)$ for all $t_m \in (0,T]$ is automatically satisfied, as can be seen directly from Equation 27. We therefore have the following proposition.

Proposition 7.6. *We have the following:*

i.

$$t_{opt}(B) = \begin{cases} 0, & \text{if } B \leq B^* \\ \min\{T, t_{max}(B)\}, & \text{if } B > B^* \end{cases}, \tag{38}$$

where

$$B^* = \min\{B > 0 : \exists t_m \in (0,T], h(t_m;B) = h(0)\}. \tag{39}$$

ii. For $B > B^*$, $t_{opt}(B)$ is a continuous function of B, and

$$\lim_{B \searrow B^*} t_{opt}(B) = t^* \in (0,T] \tag{40}$$

where $t^* = \min\{T, t_{max}(B^*)\}$. *Moreover,* $t^* = T$ *if and only if* $h'(T;B^*) \geq 0$.

As when varying λ_m, we see that as B increases, there is a discontinuous jump in the optimal amount invested in the make-or-break task at B^*, from 0 to t^*. The transition can be either to investing all of the available time in the make-or-break task (if $t^* = T$) or to investing time in both tasks (if $t^* < T$).

The proof is completely analogous to the case for λ_m, so we omit it.

7.3 Switching tasks more than once in the dynamic allocation problem does not provide any benefit

In this section, we show that allowing agents to switch tasks more than once in the dynamic allocation problem does not lead to an improvement in the expected reward of the optimal policy. More precisely, we show that by restricting ourselves to time allocation policies that either never switch tasks or start with the make-or-break task and switch only once, we can get equally high expected rewards as with unrestricted time allocation policies. Thus, given that an optimal policy exists, there will also exist an optimal policy with the specified properties (never switch or start from make-or-break and switch only once). But first, we need a rigorous definition of what a time allocation policy is. We use a rather general definition that requires the satisfaction of only a few intuitive properties.

This section relies on the theory of stochastic processes. A stochastic process is a random function of time. We also refer to the concepts of stopping time and filtration. We give a brief, intuitive description of these concepts and refer the interested reader to Bass (2011) and Karatzas and Shreve (2012) for more details.

A filtration is a technical way to describe the information known up to any specific point in time. We say that a stochastic process is "adapted to a filtration" if its value at any point in time relies only on the information known by that time, with respect to the filtration used.

A stopping time is a specific type of stochastic process which, as its name suggests, often describes the time that another process is stopped or, from another point of view, the time that an event occurs. Saying that the stopping time is adapted to the filtration means that whether or not the event occurs by some point in time follows from the information known so far, with respect to the filtration used. In our case, the event will be switching from one task to the other. Thus, the decision of whether to switch should strictly rely on information about the past performance.

We now proceed with the definition of a time allocation policy. In what follows we use superscripts m and s, instead of subscripts, to distinguish between the make-or-break and the safe-reward task.

Definition 7.7. *A time allocation policy (for the dynamic allocation problem) is a pair of stochastic processes* (τ^m, τ^s)*, defined on* $[0,T]$*, with the following properties:*

- $\tau_t^m + \tau_t^s = t$*, for all* $t \in [0,T]$
- $0 \leq \tau_{t_2}^m - \tau_{t_1}^m \leq t_2 - t_1$ *and* $0 \leq \tau_{t_2}^s - \tau_{t_1}^s \leq t_2 - t_1$*, for any* $t_1, t_2 \in [0,T]$*,* $t_1 \leq t_2$
- *For each* $t \in [0,T]$*,* τ_t^m *and* τ_t^s *are stopping times adapted to the filtration generated by* W^m.

We interpret τ_t^m and τ_t^s as the time devoted to the make-or-break task and safe-reward task, respectively, up to time t. The first condition says that the time devoted to both tasks together up to time t, is t. The second condition says that, inside an interval of time $[t_1, t_2]$, the time devoted to either task should be between 0 and $t_2 - t_1$. And the last condition makes sure that decisions on how much time to allocate to each task are based on the observed performance for the make-or-break task so far. Note that we do not allow τ_t^m and τ_t^s to depend on W^s, because the past performance in the safe-reward task has no effect on the future rewards.

Next we want to distinguish time allocation policies that switch tasks at most once and only from the make-or-break task to the safe-reward task. We call these *simple* time allocation policies. More precisely, we have the following definition.

Definition 7.8. *A time allocation policy (τ^m, τ^s) is simple if there exists some stopping time ρ adapted to the filtration generated by W^m, such that $\tau_t^m = \min\{t, \rho\}$ for each t.*

Intuitively, the above relation says that the time devoted to the make-or-break task increases linearly with time, until some point, where it stops increasing and takes the value ρ. This terminal value should depend only on the observed performance in the make-or-break task; this is the content of requiring ρ to be adapted to the filtration generated by W^m.

By using a time allocation policy (τ^m, τ^s), the reward from the safe-reward task is $v \cdot \lambda_s \cdot \tau_T^s$ and the reward from the make-or-break task is B, if $\tau_T^m \geq \Delta$, and 0 otherwise. These quantities are random, because τ^m and τ^s are themselves random. We denote probability and expectation with respect to a time allocation policy τ^m, τ^s by $\mathbb{P}^{\tau^m,\tau^s}$ and $\mathbb{E}^{\tau^m,\tau^s}$, respectively. Hence, the total expected reward associated with the time allocation policy τ^m, τ^s is

$$\mathbb{E}^{\tau^m,\tau^s}\left[v \cdot \lambda_s \cdot \tau_T^s + B \cdot \mathbf{1}_{\tau_T^m \geq \Delta}\right] = v \cdot \lambda_s \cdot \mathbb{E}^{\tau^m,\tau^s}\left[\tau_T^s\right] + B \cdot \mathbb{P}^{\tau^m,\tau^s}\left[\tau_T^m \geq \Delta\right], \tag{41}$$

where $\mathbf{1}_{\tau_T^m \geq \Delta}$ denotes the indicator function of the set $\{\tau_T^m \geq \Delta\}$.

Our central claim in this section is that instead of searching over all time allocation policies for an optimal one, it is enough to search among the simple ones. This is the content of the following proposition.

Proposition 7.9. *For any time allocation policy (τ^m, τ^s), there exists a simple time allocation policy $(\overline{\tau}^m, \overline{\tau}^s)$ with the same total expected reward.*

Proof: Let (τ^m, τ^s) be any time allocation policy and define

$$\overline{\tau}_t^m = \min\{t, \tau_T\}, \;\; \overline{\tau}_t^s = t - \overline{\tau}_t^m \tag{42}$$

It is easy to verify that $(\overline{\tau}^m, \overline{\tau}^s)$ is also a time allocation policy. Moreover, by definition, τ_T^m is a stopping time adapted to the filtration generated by W^m. Therefore, $(\overline{\tau}^m, \overline{\tau}^s)$ is simple. Finally, note that $\overline{\tau}_T^m = \tau_T^m$ and $\overline{\tau}_T^s = \tau_T^s$, so (τ^m, τ^s) and $(\overline{\tau}^m, \overline{\tau}^s)$ give the same total expected reward, by Equation 41. □

Proposition 7.9 is crucial because it reduces the dynamic time allocation problem to an optimal stopping problem, a class of stochastic optimization problems that has been extensively studied (Peskir & Shiryaev, 2006). This allows us to employ the broadly used algorithmic solution described in the next section.

7.4 Algorithmic solution for the dynamic allocation problem

The goal of this section is twofold. First, we want to describe how to numerically calculate a time allocation policy for the dynamic time allocation task that is arbitrarily close to being optimal. Second, we want to highlight the fact that calculating such a policy involves a backwards induction mechanism, which in our opinion is more readily appreciated in the discretized version of the problem, rather than in the continuous one.

The dynamic time allocation problem, in the form described in Section 3.2.1, is a stochastic optimal control problem, whose solution is described by a partial differential equation known as the Hamilton-Jacobi-Bellman equation (Bertsekas, 1995). In solving it, one has to work backwards in time, at least implicitly, since the "initial" conditions refer to the final time T. A standard method to approximate an optimal solution is to discretize the problem (Kushner & Dupuis, 2013), which leads to the discrete version of the Hamilton-Jacobi-Bellman equation, referred to as the Bellman equation (Bellman, 2013). This has the advantage of making much clearer the need for backwards induction in order to calculate the optimal policy. Here we describe the discrete version of our problem and derive the Bellman equation associated with it.

Recall that the agent has total time T to allocate between two tasks, one of which provides a reward proportional to the time invested, and the other provides a large reward B only if the performance in this task exceeds some threshold Δ. In the dynamic allocation problem, the agent at each time knows their performance so far and can use this information to adapt their strategy. The agent's performance q_m in the make-or-break task and q_s in the safe-reward task are given by

$$q_m(t_m) = \lambda_m t_m + \sigma_m W^m_{t_m} \quad \text{and} \quad q_s(t_s) = \lambda_s t_s + \sigma_s W^s_{t_s}, \tag{43}$$

respectively, where t_m is the time allocated (so far) to the make-or-break task, λ_m is the skill level parameter for this task, W^m is a Wiener process, σ_m measures the uncertainty of the performance, and similarly for the safe-reward task.

The agent then has to decide on how to distribute time between the two tasks, with the goal of maximizing the total expected reward. We consider the following discrete version of the problem: the agent may only switch tasks at times that are multiples of T/n, for some natural number n. In other words, the time T is divided into n intervals, and the agent commits to a single task during each interval. Moreover, the agent receives the reward for the make-or-break task only if their performance exceeds the reward threshold Δ at one of these discrete time points. As the number of allowed switching points increases, the difference between the discrete and continuous version of the problem becomes negligible and the expected reward of the optimal policy for the discrete version converges to the optimal reward of the continuous version (see Chapter 10 in Kushner & Dupuis, 2013).

In specifying a time allocation policy for the make-or-break task, the agent has to make n choices, at times $t_0 = 0, t_1 = \frac{T}{n}, \ldots, t_{n-1} = \frac{(n-1)T}{n}$, based on the performance for the make-or-break task at that time. As in the continuous time version of the problem, there is no benefit from starting with the safe-reward task earlier before switching to the make-or-break task; the optimal strategy involves beginning with the

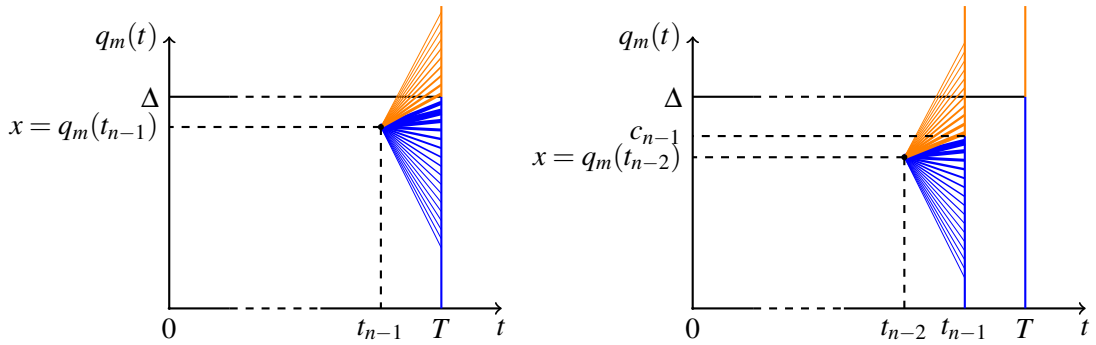

Figure 9: The process of backward induction that an agent has to follow to numerically calculate the returns from the optimal policy. **Left:** For any performance level at time t_{n-1}, the agent has to calculate the expected reward from performing the make-or-break task in the interval $[t_{n-1}, T]$, by looking at all possible terminal performance values at time T. For terminal performance values above the reward threshold Δ (denoted by orange color), the reward from the make-or-break task will be B, while for terminal performance values below Δ (blue), the reward will be 0. The expected reward of the task will be a weighted average of these two numbers. The optimal policy can be found by comparing the expected reward of the make-or-break task with that of the safe-reward task for the same interval. For large performance values at t_{n-1}, it will be optimal to invest in the make-or-break task; for smaller performance values, it will be better to invest in the safe-reward task. The optimal giving-up threshold c_{n-1} can be located by finding the break-even point where the two tasks give the same expected reward. **Right:** For any performance level at time t_{n-2}, we calculate the expected reward from performing the make-or-break task in the interval $[t_{n-2}, t_{n-1}]$ and the optimal policy in $[t_{n-1}, T]$, which is known from the previous step (orange for make-or-break task and blue for safe-reward task). We compare this expected reward with that of the safe-reward task for the whole interval $[t_{n-2}, T]$. The larger of the two will be the expected reward of the optimal policy for $[t_{n-2}, T]$. As for c_{n-1}, the optimal giving-up threshold c_{n-2} can be located by finding the point at which the agent is indifferent between the two courses of action. We continue like this for t_{n-3}, t_{n-4},...,t_0.

make-or-break task and at some point switching to the safe-reward task (Proposition 7.9). The switch may only happen at times $t_0 = 0, t_1, \ldots, t_{n-1}, t_n = T$, with t_0 corresponding to starting off with the safe-reward task and t_n corresponding to never switching. Accordingly, for any $k = 0, \ldots, n-1$, at time t_k the agent has two possible strategies: perform the safe-reward task for the rest of the time remaining; or perform the make-or-break task for one time interval and re-evaluate whether to continue or to switch tasks at time t_{k+1} (when there will be new information). The optimal choice is the one that gives a higher expected payoff.

The expected payoff on investing the remaining time in the safe-reward task can be calculated immediately. However, the payoff for the make-or-break task in the interval $[t_k, t_{k+1}]$ depends not only on the performance outcome, but also on the choices made at later times. If we know the optimal strategy to follow from time t_{k+1} onwards, we may assume that the agent will follow it. In other words, if we have already found the optimal policy in the interval $[t_{k+1}, T]$, then we may use it to calculate the payoff from choosing to perform the make-or-break task in the interval $[t_k, t_{k+1}]$. This suggests that we solve the problem with backwards induction.

The solution is illustrated in Fig. 9. We start by considering the decision at time t_{n-1}. Since there cannot be any task switching after that time, we only have to compare two strategies: perform the make-or-break task or the safe-reward task for the interval $[t_{n-1}, T]$. Recall that we are assuming that the reward threshold has not been reached, so in particular $q_m(t_{n-1}) < \Delta$. The expected reward from performing the safe-reward task is $v \cdot \lambda_s \cdot (T - t_{n-1}) = v \cdot \lambda_s \cdot \frac{T}{n}$. The reward from performing the make-or-break task depends on the

performance at time T, $q_m(T)$. If the performance is $y = q_m(T)$, the reward will be $r_m(y)$. But at time t_{n-1}, the agent does not have this information; their decision has to be based on their performance up to time t_{n-1}, $q_m(t_{n-1})$. Given a performance $x = q_m(t_{n-1})$ at time t_{n-1}, there is a probability distribution for the performance $y = q_m(T)$ at time T. Therefore, to find the expected reward at time t_{n-1}, one has to integrate over all possible performances at time T, weighted by their likelihood. This is shown in Figure 9a. In symbols, we write

$$\mathbb{E}[r_m(T)|q_m(t_{n-1}) = x] = \int_{-\infty}^{\infty} \mathbb{E}[r_m(T)|q_m(t_{n-1}) = x, q_m(T) = y] \cdot \phi_x(y)dy \tag{44}$$

$$= \int_{-\infty}^{\infty} \mathbb{E}[r_m(T)|q_m(T) = y] \cdot \phi_x(y)dy, \tag{45}$$

where $\phi_x(y)$ is the probability density function for the performance at time T, given that at time t_{n-1} the performance was x. The term on the left is the expected reward from performing the make-or-break task on the last interval, given that the performance at time t_{n-1} was x (and the reward threshold has not been reached before that). The expected value in the first integral is conditioned on the performance at time t_{n-1} being x, and the performance at time $t_n = T$ being y. But if we know the performance at the last step, the performance at earlier steps is irrelevant for calculating the reward, so this justifies the last equality.

Now, the conditional expected value in the last integral is straightforward to calculate; recall that $r_m(T) = g_m(q_m(T))$, where g_m is given by Equation 2, and we simply substitute y for $q_m(T)$. That is, Equation 45 can be rewritten as

$$\mathbb{E}[r_m(T)|q_m(t_{n-1}) = x] = \int_{-\infty}^{\infty} g_m(y) \cdot \phi_x(y)dy. \tag{46}$$

To find $\phi_x(y)$, note that the performance change for an interval of length $t_{n-1} - t_{n-2} = \frac{T}{n}$ is normally distributed, with mean $\lambda_m \cdot \frac{T}{n}$ and variance $\frac{\sigma_m^2 T^2}{n^2}$. That is,

$$\phi_x(y) = \frac{n}{\sqrt{2\pi}\sigma_m T} \cdot e^{-\frac{n^2\left(y-x-\lambda_m \frac{T}{n}\right)^2}{2\sigma_m^2 T^2}} \tag{47}$$

Equation 46 gives us the expected reward for performing the make-or-break task in the last time interval, for any observed performance x at time t_{n-1}. In order to decide which task to perform, we compare this to the expected reward of the safe-reward task, that is $v \cdot \lambda_s \cdot \frac{T}{n}$. The expected reward of the optimal policy for the last step will be

$$R_{n-1}(x) = \max\left\{\int_{-\infty}^{\infty} g_m(y) \cdot \phi_x(y)dy,\ v \cdot \lambda_s \cdot \frac{T}{n}\right\}. \tag{48}$$

For small values of x, the safe-reward task will give a higher expected reward, so that $R_{n-1}(x)$ will equal $v \cdot \lambda_s \cdot \frac{T}{n}$. Note that this term does not depend on x. For larger x (close to the reward threshold Δ), performing the make-or-break task will yield a higher expected reward, so that $R_{n-1}(x)$ will equal $\int_{-\infty}^{\infty} r_m(y) \cdot \phi_x(y)dy$, which does depend on x. We denote by c_{n-1} the break-even point, for which the expected rewards of the two tasks are equal. That is, c_{n-1} solves the equation

$$\int_{-\infty}^{\infty} g_m(y) \cdot \phi_{c_{n-1}}(y)dy = v \cdot \lambda_s \cdot \frac{T}{n}. \tag{49}$$

Once c_{n-1} has been calculated, the optimal policy for the interval $[t_{n-1}, T]$ can be simply described as follows: If $q_m(t_{n-1}) > c_{n-1}$, continue performing the make-or-break task; otherwise, switch to the safe-reward task.

Now that we know the expected reward of the optimal policy for the last step for any performance value x, we can go one step back, to calculate the expected reward at time t_{n-2}. Again, the expected reward of performing the safe-reward task is straightforward to find: $v \cdot \lambda_s \cdot (T - t_{n-2}) = v \cdot \lambda_s \cdot \frac{2T}{n}$. We now consider the expected reward for performing the make-or-break task for the interval $[t_{n-2}, t_{n-1}]$ and *assuming that the optimal policy will be followed for the interval* $[t_{n-1}, T]$, which is the case for a fully rational agent.

Suppose that at time t_{n-2} the performance is $x = q_m(t_{n-2})$. Then, there is a probability distribution for the performance $y = q_m(t_{n-1})$ at time t_{n-1}. We take this into account in calculating the expected reward for performing the make-or-break task. This is illustrated in the right part of Figure 9b, and can be expressed as

$$\mathbb{E}\left[R^m_{n-2}(x) | q_m(t_{n-2}) = x\right] = \int_{-\infty}^{\infty} \mathbb{E}\left[R^m_{n-2}(x) | q_m(t_{n-2}) = x, q_m(t_{n-1}) = y\right] \cdot \phi_x(y) dy \tag{50}$$

$$= \int_{-\infty}^{\infty} \mathbb{E}\left[R^m_{n-2}(x) | q_m(t_{n-1}) = y\right] \cdot \phi_x(y) dy, \tag{51}$$

where $R^m_{n-2}(x)$ denotes the reward of the optimal policy, if $q_m(t_{n-2}) = x$, the make-or-break task is performed in the interval $[t_{n-2}, t_{n-1}]$ and the optimal policy after t_{n-1}.

To calculate the conditional expectation in the last integral, note that since we know that $q_m(t_{n-1}) = y$ and we are assuming that the optimal policy will be followed in the last interval, the expected reward is equal to the optimal reward $R_{n-1}(y)$. Therefore, the above can be rewritten as

$$\mathbb{E}\left[R^m_{n-2}(x) | q_m(t_{n-2}) = x\right] = \int_{-\infty}^{\infty} R_{n-1}(y) \cdot \phi_x(y) dy. \tag{52}$$

The expected reward of the optimal policy is

$$R_{n-2}(x) = \max\left\{ \int_{-\infty}^{\infty} R_{n-1}(y) \cdot \phi_x(y) dy,\ v \cdot \lambda_s \cdot \frac{2T}{n} \right\}. \tag{53}$$

The above procedure can be continued inductively, to get

$$R_k(x) = \max\left\{ \int_{-\infty}^{\infty} R_{k+1}(y) \cdot \phi_x(y) dy,\ v \cdot \lambda_s \cdot \frac{(n-k)T}{n} \right\}, \tag{54}$$

for any $k = 0, \ldots, n-1$. Note that Equation 48 is a special case of Equation 54, since $R_n(y) = g_m(y)$.

Equation 54 is an instance of the Wald-Bellman equations. A more rigorous proof that the above equations give the reward of the optimal policy can be found in Section 1.2 of Peskir and Shiryaev (2006).

For any k, the performance break-even point c_k can be calculated as in Equation 55: it is the unique solution of the equation

$$\int_{-\infty}^{\infty} R_{k+1} \cdot \phi_{c_k}(y) dy = v \cdot \lambda_s \cdot \frac{(n-k)T}{n}. \tag{55}$$

For $q_m(t_k) > c_k$, the agent should continue performing the make-or-break task in the interval $[t_k, t_{k+1}]$.

Otherwise, they should switch to the safe-reward task. For consistency, we set $c_N = \Delta$.

To summarize, the algorithm for the optimal policy is the following:

- For each k and each x, find $R_k(x)$ inductively from Equation 54, and c_k from Equation 55. Initialize with $R_N(x) = g_m(x)$, where g_m is given by Equation 2, and $c_N = \Delta$.
- If $c_0 \geq 0$, then perform the safe-reward task only.
- If $c_0 < 0$, start with the make-or-break task and switch to the safe-reward task at time

$$\tau = \min\{k : q_m(t_k) \leq c_k \text{ or } q_m(t_k) \geq \Delta\}, \tag{56}$$

  with the understanding that $\tau = t_N$ implies that there is no switching.

7.5 Analytic expressions of hitting times and expected returns for the play-to-win strategy

In this section we provide a formula for calculating the expected reward for the play-to-win strategy. Recall from Section 3.2.2 that the time τ at which an agent using the play-to-win strategy will switch to the make-or-break task is given by

$$\tau = \min\{t : q_m(t) \geq \Delta\} = \min\left\{t : W_t^m + \frac{\lambda_m}{\sigma_m} \cdot t \geq \frac{\Delta}{\sigma_m}\right\}, \tag{57}$$

as long as $\tau < T$. The expected reward from the make-or-break task is then

$$\mathbb{E}[r_m(t_m)] = B \cdot \mathbb{P}(\tau \leq T) \tag{58}$$

and the expected reward from the safe-reward task is

$$\begin{aligned}
\mathbb{E}[r_s(t_s)] &= \mathbb{E}[v \cdot \lambda_s \cdot (T - \tau) \cdot \mathbf{1}_{\tau \leq T}] && (59)\\
&= \mathbb{E}[v \cdot \lambda_s \cdot T \cdot \mathbf{1}_{\tau \leq T}] - \mathbb{E}[v \cdot \lambda_s \cdot \tau \cdot \mathbf{1}_{\tau \leq T}] && (60)\\
&= v \cdot \lambda_s \cdot T \cdot \mathbb{P}(\tau \leq T) - v \cdot \lambda_s \cdot \mathbb{E}[\tau \cdot \mathbf{1}_{\tau \leq T}], && (61)
\end{aligned}$$

where $\mathbf{1}_{\tau \leq T}$ denotes the indicator function of the set $\{\tau \leq T\}$; it equals 1 if $\tau \leq T$ and 0 otherwise.

Therefore, the total expected reward for the play-to-win strategy is

$$\mathbb{E}[r_m(t_m) + r_s(t_s)] = (v \cdot \lambda_s \cdot T + B) \cdot \mathbb{P}(\tau \leq T) - v \cdot \lambda_s \cdot \mathbb{E}[\tau \cdot \mathbf{1}_{\tau \leq T}]. \tag{62}$$

To continue, we need an expression for the probability density function of τ. From Equation 57 we see that τ is the hitting time at level $\frac{\Delta}{\sigma_m}$ of a Brownian motion with drift $\frac{\lambda_m}{\sigma_m}$ (Bass, 2011; Karatzas & Shreve, 2012). Its probability density is an inverse Gaussian distribution (see Section 3.2 in Chhikara, 1989), given for any $t > 0$ by

$$f_\tau(t) = \frac{\Delta}{\sqrt{2\pi t^3} \cdot \sigma_m} \cdot e^{-\frac{(\Delta - \lambda_m \cdot t)^2}{2t{\sigma_m}^2}}. \tag{63}$$

Using this, we can write the total expected reward as

$$\mathbb{E}[r_m(t_m) + r_s(t_s)] = \frac{\Delta}{\sqrt{2\pi} \cdot \sigma_m} \cdot \int_0^T \left[(v \cdot \lambda_s \cdot T + B) t^{-\frac{3}{2}} - v \cdot \lambda_s \cdot t^{-\frac{1}{2}} \right] \cdot e^{-\frac{(\Delta - \lambda_m \cdot t)^2}{2t{\sigma_m}^2}} dt. \tag{64}$$