© 2019. Published by The Company of Biologists Ltd | Journal of Experimental Biology (2019) 222, jeb198234. doi:10.1242/jeb.198234

RESEARCH ARTICLE

How humans initiate energy optimization and converge on their optimal gaits

Jessica C. Selinger[1,2,*], Jeremy D. Wong[2,3], Surabhi N. Simha[2] and J. Maxwell Donelan[2]

ABSTRACT

A central principle in motor control is that the coordination strategies learned by our nervous system are often optimal. Here, we combined human experiments with computational reinforcement learning models to study how the nervous system navigates possible movements to arrive at an optimal coordination. Our experiments used robotic exoskeletons to reshape the relationship between how participants walk and how much energy they consume. We found that while some participants used their relatively high natural gait variability to explore the new energetic landscape and spontaneously initiate energy optimization, most participants preferred to exploit their originally preferred, but now suboptimal, gait. We could nevertheless reliably initiate optimization in these exploiters by providing them with the experience of lower cost gaits, suggesting that the nervous system benefits from cues about the relevant dimensions along which to re-optimize its coordination. Once optimization was initiated, we found that the nervous system employed a local search process to converge on the new optimum gait over tens of seconds. Once optimization was completed, the nervous system learned to predict this new optimal gait and rapidly returned to it within a few steps if perturbed away. We then used our data to develop reinforcement learning models that can predict experimental behaviours, and applied these models to inductively reason about how the nervous system optimizes coordination. We conclude that the nervous system optimizes for energy using a prediction of the optimal gait, and then refines this prediction with the cost of each new walking step.

KEY WORDS: Sensorimotor control, Exoskeletons, Metabolic cost, Energy expenditure, Reinforcement learning

INTRODUCTION

People often learn optimal coordination strategies. That is, the nervous system has an objective for movement and it adapts its coordination to minimize this objective function. This optimization principle underlies theories on the control of reaching, grasping, gaze and gait, although the nervous system may seek to minimize different objectives for each of these tasks (Alexander, 1996; Flash and Hogan, 1985; Kording et al., 2007; Kuo and Donelan, 2010; Scott and Norman, 2003; Shadmehr et al., 2016; Srinivasan and Ruina, 2005; Todorov, 2004; Todorov and Jordan, 2002). This principle has provided insight into healthy and pathological behaviour, as well as the functions of different brain areas (Bastian, 2008; Krakauer, 2006; Shadmehr and Krakauer, 2008; Wolpert et al., 2011). While there is a growing body of evidence that preferred behaviour in these various tasks indeed optimizes reasonable objective functions, how the nervous system performs this optimization is largely unknown (Franklin and Wolpert, 2011; Todorov, 2004). This is because the primary focus of past work has been on identifying the objectives that explain our existing behaviours, rather than on how these objective functions are minimized over time (Franklin and Wolpert, 2011; Todorov, 2004). Here, we used both experiments and models to understand how the nervous system learns to optimize our movements.

The optimization of movement is a challenge for the nervous system. To perform a movement, the nervous system has thousands of motor units at its disposal, and it can finely vary each motor unit's activity many times per second. This flexibility results in a combinatorially huge number of candidate control strategies for performing most movements – far too many for the nervous system to simply try each one to evaluate its objectives (Bellman, 1952; Bernstein, 1967). The nervous system must instead efficiently search through its options to seek optimal solutions within usefully short periods of time. A second consequence of the large number of control strategies available to the nervous system is that it can never know whether it has truly converged to the best of all possible options. But if it is indeed at an optimum, continuously searching for better options will itself be suboptimal (Sutton and Barto, 1998). Thus, the nervous system must determine when to initiate optimization and explore new coordination patterns, and when to exploit previously learned strategies (Tumer and Brainard, 2007; Wilson et al., 2014; Wu et al., 2014).

Here, we used human walking to understand how the nervous system initiates and performs the optimization of its motor control strategies. Human walking is a system well suited for studying these questions because the primary contributor to the nervous system's objective function – metabolic energy expenditure – is both well established and directly measurable. Decades of experiments using respiratory gas analysis have established that our preferred gait parameters – from walking speed to step frequency and step width – minimize energetic cost (Atzler and Herbst, 1928; Donelan et al., 2001; Elftman, 1966; Minetti et al., 1993; Molen et al., 1972; Ralston, 1958; Umberger and Martin, 2007; Zarrugh et al., 1974). While some optimal motor control strategies may be established over relatively long periods of time, we recently discovered that the nervous system can re-optimize aspects of gait within minutes (Selinger et al., 2015). This allows us to observe energy optimization within a lab setting and within a reasonably short period of time. Studying optimization in tasks such as reaching or saccades is less straightforward as the nervous system's objective function appears to include a term not only for task effort but also for task error, with some unknown weighting between these two contributors (Scott, 2004; Shadmehr et al., 2016; Todorov, 2004;

[1]School of Kinesiology and Health Studies, Queen's University, Kingston, ON, Canada, K7L 3N6. [2]Department of Biomedical Physiology and Kinesiology, Simon Fraser University, Burnaby, BC, Canada, V5A 1S6. [3]Faculty of Kinesiology, University of Calgary, Calgary, AB, Canada, T2N 1N4.

*Author for correspondence (j.selinger@queensu.ca)

J.C.S., 0000-0001-9372-6705

Received 27 December 2018; Accepted 2 September 2019

Wolpert and Ghahramani, 2000). Furthermore, motor learning in these tasks appears, at least initially, to prioritize reducing error over optimizing energy cost, requiring creative experiments to decouple error-based learning from reward-based learning (Krakauer and Mazzoni, 2011; Wolpert et al., 2011).

To study how the nervous system performs energy optimization in human walking, we leveraged our previously developed experimental design within which people reliably optimize their gait to minimize energetic cost (Selinger et al., 2015). This design used robotic exoskeletons to reshape the relationship between step frequency and energetic cost – which we term the cost landscape – shifting the energetically optimal step frequency away from the normally preferred and optimal step frequency. When given sufficient experience with the new cost landscape, participants in our past experiments learned to adapt their step frequency to converge on the new energetic minimum (Selinger et al., 2015). More recently, we have found the same to be true for adapting step width, and when using other methods for changing cost landscapes (Abram et al., 2019; Simha et al., 2019). We use the term optimization to refer to the process of adapting coordination towards new patterns that minimize the objective function (in our case energy cost). This might alternatively be called reward-based adaptation (Krakauer and Mazzoni, 2011; Wolpert et al., 2011). We also distinguish between optimization and prediction, where the former is the process of trying new coordination patterns as the nervous system converges towards the minimum, and the latter is the nervous system storing and recalling previously experienced coordination patterns (O'Connor and Donelan, 2012; Pagliara et al., 2014). For our purposes, we consider prediction, because it involves the storage and recall of a coordination pattern, as commensurate with learning.

While our prior work demonstrated that the nervous system can continuously optimize energetic cost, it did not allow us to decipher the nervous system's mechanisms for this optimization. To understand these mechanisms, here we used a series of experiments that controlled the type of initial experience participants received with a new energetic cost landscape to determine what gait experience was sufficient for the nervous system to stop exploiting a previously optimal solution and initiate a new optimization. We designed the new experiments to isolate the different types of experience participants received during the exploration period of our previous experiment, which was broad, varied and sufficient to elicit optimization (Selinger et al., 2015). We considered three possibilities: (1) the nervous system can spontaneously initiate optimization, (2) the nervous system can initiate optimization after experience with discrete points on the new landscape, and (3) the nervous system only initiates optimization in response to broad experience with the new landscape. Once the nervous system initiated optimization, we studied how it explored new gaits, in order to understand the nervous system's algorithms for converging on new energetic optima. Based on common algorithms in numerical optimization, we again considered three possibilities: (1) the nervous system uses a 'choose best' strategy where it remains at the gait with the lowest experienced cost, (2) the nervous system uses a 'sampling' strategy where it intermittently explores a range of gaits, or (3) the nervous system uses a 'local search' strategy where it continues to adjust a given gait parameter as long it continues to result in cost reductions.

In addition to experiments, we use computational models of the nervous system's optimization to understand how the nervous system initiates and converges on optimal movements. These models use reinforcement learning algorithms to iteratively learn and then rapidly predict the energy optimal gait (Sutton and Barto, 1998; Sutton et al., 1992). We chose a reinforcement learning approach for two reasons. First, it has been used to successfully find the optimal coordination for walking and reaching in robots and physics-based simulations (Collins et al., 2005; Lillicrap et al., 2016 preprint; Peters and Schaal, 2008). We view this as a proof of principle for human motor control. Second, the necessary physiological components for humans to perform reinforcement learning, including reward prediction and sensory feedback, are present in our nervous systems and well studied for learning non-motor tasks (Schultz et al., 1997). We used the findings from our specific experiments to guide the development of these reinforcement learning models, and the models to inductively reason about how the nervous system optimizes coordination.

MATERIALS AND METHODS

Experiments

Participants

We performed testing on a total of 36 healthy adults (body mass: 63.9±9.8 kg; height: 1.69±0.10 cm; means±s.d.) with no known gait or cardiopulmonary impairments. Simon Fraser University's Office of Research Ethics approved the protocol, and participants gave their written, informed consent before experimentation.

Exoskeleton hardware and controller

We manipulated energetic cost using robotic exoskeletons mounted about the knee joints (Fig. 1A,B) (Selinger et al., 2015). Each exoskeleton weighed 1.1 kg and was composed of a custom carbon fibre shell and custom steel gear train coupled to an off-the-shelf rotary magnetic motor (BLDC40S-10A, NMB Technologies Inc.). During walking, the relatively low angular velocity characteristic of knee motion was transformed by the gear train to produce relatively high angular velocity at the motor. This rotational motion in the motor's rotor induced voltage in the motor's windings and, when allowed, electrical current. The induced current generated its own magnetic field that resisted the motion of the knee with a torque proportional to the current magnitude. We used a custom control unit to measure and control the flow of electrical current through the motor, and therefore the magnitude of the resistive torque applied to the knees.

All participants experienced a 'penalize-high' control function (Fig. 1C–G), where the applied resistance, and therefore added energetic penalty, was minimal at low step frequencies and increased as step frequency increased (Selinger et al., 2015). Our past experiments demonstrated that this control function reshapes the relationship between step frequency and energetic cost, creating a positively sloped energetic gradient at the participant's initial preferred step frequency, and an energetic minimum at a lower step frequency. To implement this controller, we made the commanded resistive torque sent to the control unit proportional to the participant's step frequency measured from the previous step. We calculated the step frequency for an individual step as the inverse of the time between foot contact events, identified from the characteristic rapid fore–aft translation in ground reaction force centre of pressure from the instrumented treadmill (FIT, Bertec Inc.). We sampled ground reaction forces, and measured motor current and voltage, at 200 Hz (NI DAQ PC1-6071E, National Instruments Corporation). We commanded step frequency, and the newly desired knee torque, in real time at 200 Hz using custom software (Simulink Real-Time Workshop, MathWorks). In the controller off setting, the commanded current, and thus commanded resistive torque, was zero.

Experimental protocol

The protocol consisted of four distinct periods: Baseline Period, Habituation Period, First Experience Period, and one of the three possible Second Experience Periods (Fig. 2A–F, respectively,

Journal of Experimental Biology (2019) 222, jeb198234. doi:10.1242/jeb.198234

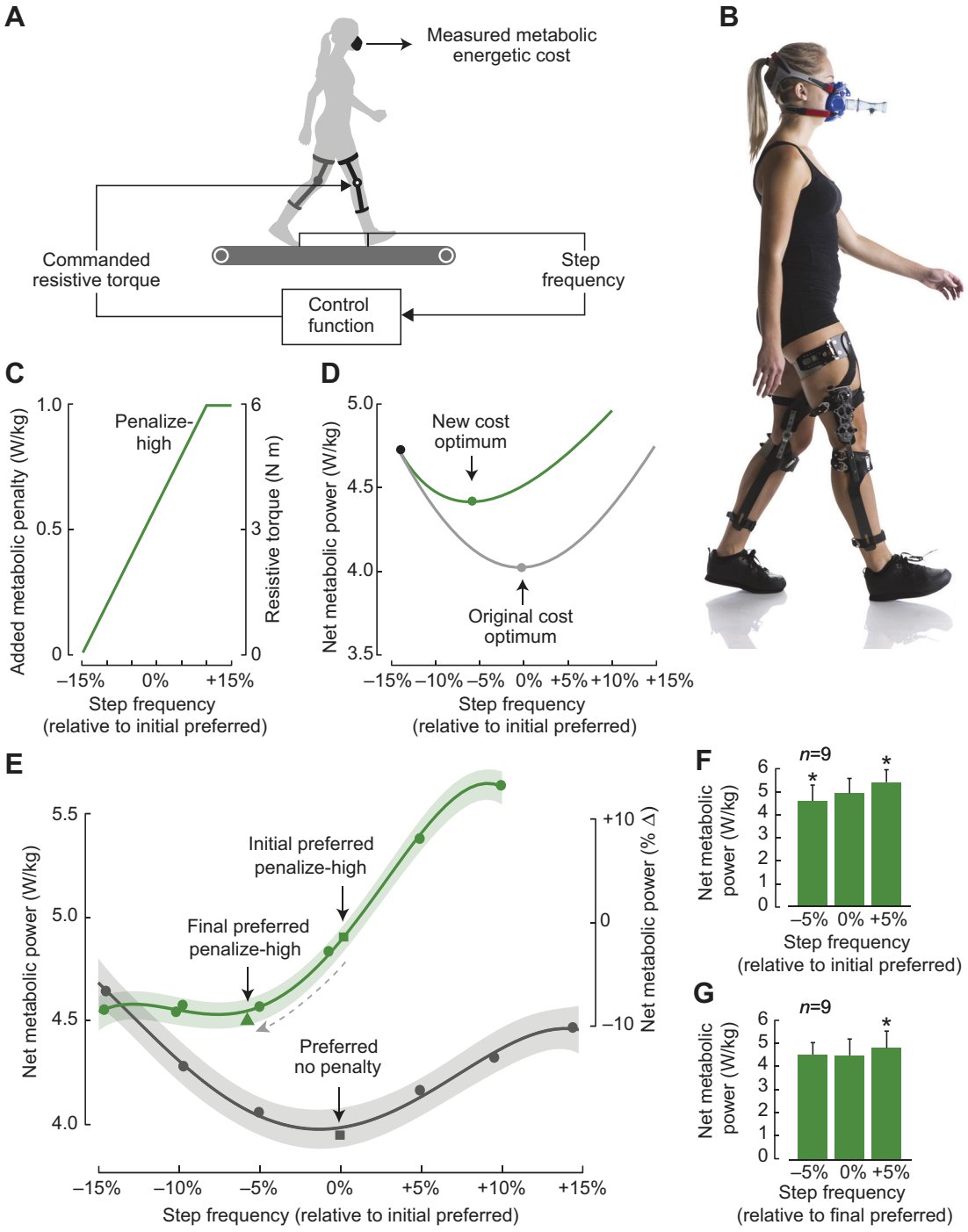

Fig. 1. Experimental design. (A,B) By controlling a motor attached to the gear train of our exoskeletons, we can apply a resistance to the limb that is proportional to the participant's step frequency. (C) Design of the penalize-high (green) control function. (D) Schematic energetic cost landscapes. Adding the energetic cost of the penalize-high control function to the original cost landscape (grey) produces a new cost landscape with the optimum shifted to lower step frequencies (green curve). (E) Measured energetic cost landscapes (reproduced from Selinger et al., 2015) for the penalize-high (green) control function and controller off condition (grey). The lines are 4th order polynomial fits with 95% confidence intervals (shading), shown only for illustrative purposes. The dashed grey arrow illustrates the direction of adaptation from initial preferred (green square) to final preferred step frequencies (green triangle). On average, participants decreased their step frequency by approximately 6% to converge on the energetic minima and reduce cost by 8%. (F) The penalize-high control function creates a positively sloped energetic gradient about the participants' initial preferred step frequency. (G) Participants adapted their step frequency to converge on the energetic minima. Error bars represent 1 s.d. Asterisks indicate statistically significant differences in energetic cost when compared with the cost at the initial or final preferred step frequency (0%).

described in detail below). We used the Baseline Period to identify participants' normally preferred gait and the Habituation Period to familiarize participants with walking at a range of gaits. Participants experienced the new cost landscape for the first time during the First Experience Period, allowing us to determine whether they spontaneously optimized their gait. These three periods of the experimental protocol were common to all participants. During the Second Experience Period, we systematically varied the type of

Journal of Experimental Biology (2019) 222, jeb198234. doi:10.1242/jeb.198234

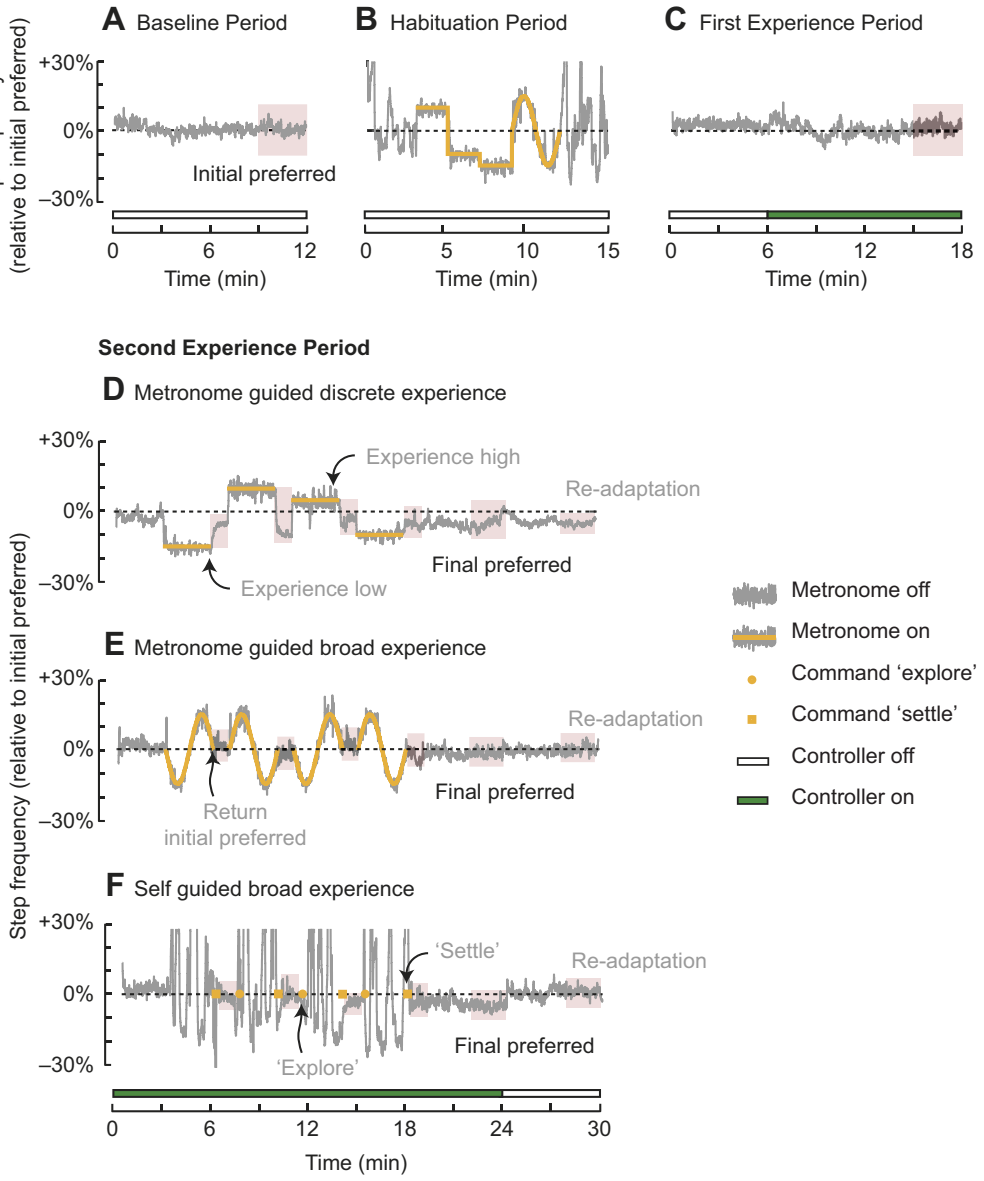

Fig. 2. Experimental protocol. Each participant completed four testing periods. The first three – a Baseline Period (A), Habituation Period (B) and First Experience Period (C) – were the same for all participants. For the final one, participants were assigned to one of three possible Second Experience Periods: a metronome guided discrete experience (D), metronome guided broad experience (E) or self guided broad experience (F). Rest periods of 5–10 min were provided between each testing period. For all periods, regions of red shading illustrate the time windows during which we assessed steady-state step frequency. Data shown in A–C and E are from one representative participant, while data in D and F are from two other representative participants.

experience that participants received with the new cost landscape to study which types were sufficient to initiate optimization. To accomplish this, we randomly assigned participants to one of the three groups that received different kinds of experience during this period. All four periods were completed in one testing session, which lasted 3 h and had no more than 2.5 h of walking to reduce fatigue effects. Participants were given between 5 and 10 min of rest in between each of the four walking periods. Table S1 provides a detailed outline of initial, added and final participant numbers for each experimental condition.

At the beginning of testing, we instrumented participants with the exoskeletons and indirect calorimetry equipment (VMax Encore Metabolic Cart, VIASYS®). We then determined their resting energetic cost during a 6 min quiet standing period. Following this, all participants completed the Baseline Period of 12 min of walking while wearing the exoskeletons, but with the controller turned off (Fig. 2A). The exoskeleton added small inertial and frictional torques when the controller was off, but the motor added no additional resistance (Selinger et al., 2015). We used the average of the final 3 min of walking data to determine participants' 'initial preferred step frequency'.

Next, participants completed a Habituation Period where they were familiarized with walking at a range of step frequencies (Fig. 2B). During this Habituation Period, the controller remained off so that participants did not gain experience with the new energetic landscape. We first encouraged participants to explore walking with very long slow steps or very short fast steps. Participants next practiced walking at three different steady-state tempos, played for 3 min each. These tempos were +10%, −10% and −15% of the initial preferred step frequency. In the final stage of habituation, participants practiced walking to a sinusoidally varying metronome tempo, which had a range of ±15% of the initial preferred step frequency and a period of 3 min.

Following the Habituation Period, we explained to participants that they would next walk for 6 min with the exoskeleton turned off, at which point the exoskeleton would turn on and they would walk for a further 12 min. They were given no other directives about how to walk and at no point during testing were participants provided with any information about how the controller functioned, or how their step frequency influenced the resistance applied to the limb. The participants then completed the First Experience Period,

beginning with 6 min of walking with the controller off, followed by 12 min of walking with the controller on (Fig. 2C).

Next, each participant completed one of the three Second Experience Periods (described below). Prior to beginning this testing period, we informed all participants that they would be walking for 30 min and that the exoskeleton controller would be on for the first 24 min and off for the final 6 min. To gain insight into the progress of optimization during each Second Experience Period, 1 min probes of participants' self-selected step frequency occurred at the 6th, 10th and 14th minute, along with a final 6 min probe at the 18th minute (Fig. 2D–F).

Metronome guided discrete experience

We informed participants assigned to this experience that at times the metronome would be turned on, during which they should match their steps to the steady-state tempo, and that when the metronome turned off, they no longer had to remain at that tempo (Fig. 2D). We gave participants no further directives about how to walk. The metronome tempos were −15%, −10%, +5% and +10% of initial preferred step frequency and we played each for 3 min. We chose these tempos such that they spanned the energetic landscape but did not include step frequencies explicitly at the expected optima or the preferred step frequency (approximately −5% and 0%, respectively). We randomized the order of the tempos. We turned off the metronome following each tempo for the 1 min probes of participants' self-selected step frequency.

Metronome guided broad experience

We provided those assigned to this experience with the same instructions as those in the metronome guided discrete experience, except that in this case the metronome tempo would change slowly over time (Fig. 2E). A sinusoidally varying metronome tempo was played for 3 min, four separate times, which were once again separated by 1 min probes of self-selected step frequency. The sinusoidal tempo had a range of ±15% of the initial preferred step frequency, a period of 3 min, and began and ended at 0% of the initial preferred step frequency. In this manner, the metronome guided participants through the complete landscape but always returned them to their preferred step frequency prior to a probe.

Self guided broad experience

We informed those assigned to this experience that at times we would verbally give them the command 'explore', at which point they should explore walking at a range of different step frequencies, and that they should continue to do so until we give them the command 'settle', at which point they should settle into a steady step frequency (Fig. 2E). We gave them no directives about what their steady-state step frequency should be. We instructed participants to 'explore' four separate times, each lasting 3 min and once again separated by 1 min probes of self-selected step frequency. When we gave the command 'settle', participants could be at any self-selected step frequency.

Experimental outcome measures

As defined earlier, we calculated each participant's initial preferred step frequency as the average step frequency during the final 3 min of the Baseline Period. Individual participant's variability in step frequency, calculated as a coefficient of variation, was also assessed during this time period. To asses participants' optimization progress when first exposed to the new cost landscape, average step frequency was calculated during the final 3 min of the First Experience Period. To assess the progress of optimization throughout the Second Experience Period, we calculated average step frequency during the final 30 s of each probe. To assess participants' level of optimization, following all provided experiences, we calculated the 'final preferred step frequency' as the average step frequency during the 21st to 24th minute of the Second Experience Period, just prior to the controller being turned off. To assess whether participants re-optimized when returned to the natural landscape, the 're-adaptation step frequency' was calculated as the average step frequency during the final 3 min of the Second Experience Period, when the controller was turned off. To determine whether average step frequency values were different from initial preferred step frequency values (0%), we used *t*-tests with a significance level of 0.05. To determine whether average step frequency values were different from the optimal step frequency determined from our prior experiment (5.7 ±3.9%; Selinger et al., 2015), we used two-sample *t*-tests with a significance level of 0.05.

To understand how participants were converging to optimal gaits, we also assessed the rate at which participants adapted their step frequency throughout the protocol. To assess adaptation rates when first exposed to the new landscape, we used step frequency time series data from minutes 6 to 18 of the First Experience Period, when the controller was first turned on. To assess the progress of optimization during the Second Experience Period, we used data from the 1 min probes. To asses re-adaptation, we also used data from minutes 24 to 30 of the Second Experience Period, when the controller was turned off. In all cases, step frequency time series data were grouped across participants of the same protocol and fitted with a single-term time-delayed exponential. For plotting purposes, we also averaged across participants in the same protocol and calculated the across-participant standard deviation at each time point.

Identifying spontaneous initiators

A subset of participants displayed gradual adaptations in gait during the First Experience Period and converged to lower, less costly, step frequencies consistent with the energetic optima. We labelled these participants as 'spontaneous initiators' if they met two criteria. First, during the final 3 min of the First Experience Period, we required their average step frequency be below 3 s.d. in steady-state variability, determined from the final 3 min of the Baseline Period. For most participants, this equated to a decrease in step frequency of approximately 5%. Second, the change in step frequency could not be an immediate and sustained response to the exoskeleton turning on. Such a response implies a mechanical or reflex response to a change in knee resistance, rather than optimization (Lam et al., 2006). Participants' final step frequency at the end of the First Experience Period had to be significantly lower than the step frequency measured in the 10th to 40th second after the exoskeleton turned on (one-tailed *t*-test, $P<0.05$). Fig. S1 illustrates how we discriminated spontaneous initiators from those who did not meet these criteria, whom we term 'non-spontaneous initiators'. We used a two-sample *t*-test with a significance level of 0.05 to compare average step frequency during the final 3 min of the First Experience Period for the spontaneous initiators and non-spontaneous initiators.

We analysed spontaneous initiators as a new group, separate from the three second experience groups. We reasoned that including them would have obscured insight into the effects of the additional experience because they had already converged to their new optima. We added additional participants to the second experience groups to rebalance our conditions (see Table S1). We tested whether high natural gait variability, which results in a more expansive and therefore more clear sampling of the new cost landscape, may be a predicator of spontaneous initiation. To do so, we used a one-tailed two-sample *t*-test with a significance level of 0.05 to compare step-

to-step variability prior to the controller even being turned on (during the final 3 min of the Baseline Period) for the spontaneous initiators and non-spontaneous initiators.

Identifying effects of high and low cost experience

Our initial analysis of data from the Second Experience Periods indicated that the first experience with low step frequencies, and therefore low costs, had a lasting effect on the participants' self-selected step frequency. We used two-sample *t*-tests with a significance level of 0.05 to compare self-selected step frequencies during the first probe, following experience with high or low costs, for the metronome guided discrete experience group and the self guided broad experience group.

To investigate the different effects of high and low cost experience, along with the order of the experience, we added additional participants to the metronome guided discrete experience group (see Table S1). For the added participants, the experience prior to the first or last probe was set to be either the highest (+10%) or lowest (−15%) step frequency, with all other step frequencies assigned in random order. In total for the analysis, five participants experienced +10% and four experienced −15% prior to the first probe. Prior to the last probe, four participants experienced +10% and five experienced −15%. While these participant numbers are low, to detect an across-participant average difference in step frequency of at least 5%, given an across-participant average standard deviation in step frequency of 2.5%, we calculated that we required only four participants per group to achieve a power of 0.8. To determine the effects of high and low cost experience, as well as their order, we used a two-way ANOVA to compare the preferred step frequencies during the first and last probes following experience with the highest and lowest step frequencies. We then used *post hoc* two-sample *t*-tests with a significance level of 0.05 to compare effects of high and low cost experience during the first and last probes.

Modelling

Simple reinforcement learning model

We first tested whether a 'simple reinforcement learning model' can reproduce the experimental behaviours observed during energy optimization. Reinforcement learning, applied to our context, allows the nervous system to iteratively learn a 'value function' (Q) that is the predicted relationship between step frequency and energetic cost (i.e. a predicted cost landscape). For each new step, the nervous system selects a step frequency, or 'action' (a), in accordance with its 'policy' (π), which is to choose the energy minimal step frequency. We use a_i to represent one executed step frequency. Each time the nervous system executes a new step frequency, it measures the resulting energetic cost, or 'reward' (r), and updates its predicted cost for that step frequency. Here, we make the assumption that the reward cannot be measured perfectly because of 'measurement noise' (n_m); nor can the action be executed perfectly because of 'execution noise' (n_e). Consequently, the nervous system does not simply replace the old predicted value with the new reward. Instead, it updates the old value by some fraction of the measured reward, referred to as the 'learning rate' (α), according to the equation:

$$Q(a_i) = Q(a_i) + \alpha(r - Q(a_i)). \quad (1)$$

Fig. 6A summarizes this reinforcement learning algorithm. As a compromise between clarity and generality, we elected to use conventional reinforcement learning terminology (e.g. value-function) and naming conventions (e.g. Q), and where sensible indicate what these terms and names represent in our particular experiment (e.g. predicted cost landscape).

The learner learns a value-function (predicted cost landscape) that is a prediction of the actual value returned by the environment (actual cost landscape). We refer to the predicted cost landscape as Q, and the actual cost landscape as Q^*. In our experiments, the actual cost landscape is initially the original cost landscape – the dependence of energetic cost on step frequency during natural walking when the controller is turned off. We modelled it as the following quadratic function:

$$Q^*_{off}(A) = 10 \cdot \left(\frac{A}{100}\right)^2 + 1, \quad (2)$$

where A is a set of 35 possible step frequencies, or actions, that range between −17% and +17%. Our choice to discretize the action space enforces local learning, where actions at distinct step frequencies have no effect on the expected value of others. It is possible, if not likely, that the nervous system does not discretize its action space in this way but may rather store a representative function. While the choice to discretize the action space may affect the time course of individual simulations, the general behavioural features of energy optimization we find are unaffected. The predicted cost has a normalized cost of one at the optimum and a curvature that approximates our experimentally measured landscape (Fig. 6C). To simulate the controller turning on, we changed Q^* to the new cost landscape – the dependence of energetic cost on step frequency during walking under our control function. We modelled it as:

$$Q^*_{on}(A) = Q^*_{off}(A) + \left(\frac{A}{60} + \frac{1}{4}\right), \quad (3)$$

where the cost added to Q^*_{off} approximates the energetic effect of our controller, creating a landscape similar in shape to that which we measure experimentally (Fig. 6C).

Parameter sensitivity analysis

We performed a sensitivity analysis to determine feasible ranges for model parameters that are consistent with both experimentally measured rates of convergence to the optimum and experimentally measured variability in step frequency. To do so, we repeatedly simulated a protocol that was similar in design to our experimental First Experience Period. The simulated protocol lasted 1440 walking steps (approximately 12 min) in which the landscape changed from Q^*_{off} to Q^*_{on} after 720 steps (approximately 6 min). For each step, the simple reinforcement learning model chose its step frequency by applying its policy to its current value function, and then updated its value function with each new reward. Also, at each step, the contribution of measurement noise to the sensed reward (n_m) was sampled from a Gaussian distribution with zero mean and a non-zero standard deviation. We explored standard deviations that ranged between 0.1% and 6.0% of the energetic cost at the initial preferred step frequency during natural walking. We modelled the contribution of execution noise to the executed action (n_e) in the same manner, exploring values that ranged between 1.0% and 3.0% of the initial preferred step frequency. The contribution of a measured reward to the update of the value function is determined by the learning rate (α) – we explored learning rates that ranged between 0.01 and 1.00.

Because of the stochastic nature of the measurement and execution noise, we repeated simulations 1000 times for each combination of parameter settings. We then determined the rate of

convergence to the optimum by averaging simulated step frequency data across the repeated simulations and then fitting the final 720 steps with a single-process exponential model. Higher learning rates, which put greater weight on new measurements as opposed to past measurements, led to faster convergence to the optimum (Fig. S2A). This rate of convergence was largely unaffected by measurement noise, and only minimally affected by execution noise, where higher execution noise slowed convergence to the optimum. In our experiments, the convergence to the new optimum frequency typically occurred with a time constant of about 100 steps. Constraining our simulations to perform with similar rates of convergence yielded a wide range of possible learning rate parameter settings, from 0.5 to 1.0 for any of our combinations of measurement and execution noise. For the remaining simulations, we used a learning rate of 0.5.

Given this learning rate, we next selected measurement and execution noise levels that generated steady-state variability in step frequency that well approximated our experimental observations (1.0% to 1.5% standard deviations as a percentage of the mean step frequency). For each simulation repeat, we calculated the standard deviation in step frequency during the first 720 steps. During this time, the learner is at the Q^*_{off} optimum and the landscape is unchanging, resulting in steady-state behaviour. We then averaged this value across repeats to get an average measure of variability in steady-state step frequency for each of our combinations of measurement noise and execution noise. Once again, our experimental constraints left us with a wide range of possible parameter settings (Fig. S2B). For the purposes of our simulations, we set the measurement noise to be 2.0% and the execution noise to be 1.5%. Within the ranges deemed reasonable by our experimental constraints, the qualitative behaviours generated by our model are not particularly sensitive to the specific learning rate, measurement noise and execution noise parameter settings.

Reference cost reinforcement learning model

We also tested a reinforcement learner that prioritizes the learning of a 'reference cost', defined as the cost at the predicted optimum step frequency (Adams, 1971, 1976; Wolpert and Miall, 1996). With each step, this model continuously relearns the value of the reference cost and then shifts the costs associated with all frequencies by this value (Fig. 6B). It is initially free to self-select its step frequency and thus executes an action at or near the expected optima. If the action (a_i) is at the predicted optimum step frequency, it receives a new reward (r) and uses this to update the offset of the entire value function according to the equation:

$$Q(A) = Q(A) + \alpha(r - Q(a_i)). \quad (4)$$

For example, if the current predicted cost landscape is bowl shaped, and the cost of a new step is higher than the minimum value of this bowl, the updated predicted cost landscape would remain bowl shaped, but would be shifted upwards to reflect the newly experienced cost. If the action is away from the predicted optimum step frequency, the model does not update the value function. The algorithm proceeds like this until it detects a cost saving with respect to this continuously updated reference cost, after which it initiates optimization. The learner then proceeds identically to the simple reinforcement learning model above, updating the cost associated with the individual frequencies that it executes, thereby learning the shape of the entire cost landscape and not just the value of the reference cost. It is unclear from our experiments exactly what constitutes sufficient experience with a low cost gait to initiate optimization. In keeping with our experimental findings, here we assume that the criteria have been met during the metronome guided experience with low cost step frequencies prior to the first probe.

Energetically optimal learning rates

Principles of energetic optimality may also determine the choice of learning rate. It is possible to solve for a learning rate that minimizes energy expenditure; however, the optimal learning rate is dependent on how frequently the energetic landscape is changing. To demonstrate this, we simulated protocols where the landscape alternates between Q^*_{off} and Q^*_{on} with a period varying between 1 min and 12 h, and a duty cycle of 50%. We simulated 24 h of walking and evaluated learning rates ranging between 0.01 and 1.00. We kept the measurement and execution noise constant at their nominal values of 2.0% and 1.5%, respectively. We repeated model simulations 100 times for each period of landscape change and each learning rate. We then determined the average energetic cost across all steps (before measurement noise was applied), and then averaged across repeats to get an average energetic cost for each combination of period and learning rate. Finally, we solved for the learning rate that minimized the energetic cost for each period.

RESULTS

High natural gait variability can spontaneously initiate optimization

During this First Experience Period, we identified six of the 36 participants to be spontaneous initiators. They displayed gradual adaptations in gait to converge to lower, less costly, step frequencies consistent with the energetic optima in the new cost landscape (Fig. 3A,B). On average, the spontaneous initiators converged toward the optima with an average time constant of 65.7±2.7 s, or about 120 steps. They settled on step frequencies significantly lower than their initial preferred values (−8.0±2.5%, P=0.0005; Fig. 3D). These new frequencies were not different from the optimal step frequency measured in our previous experiment (P=0.23; Selinger et al., 2015). The non-spontaneous initiators remained at step frequencies that were not different from their initial preferred step frequency (0.8±2.7%; Fig. 3D).

Prior to the controller being turned on, these spontaneous initiators displayed higher variability in step frequency than non-spontaneous initiators (1.5±0.3% and 1.1±0.3%, respectively, P=0.018; Fig. 3C). This suggests that high natural gait variability is a predictor of spontaneous initiation. In support of this, we found a weak but significant correlation (R^2=0.22, P=0.004) across all 36 participants between individual participant's step frequency adaptation during the First Experience Period and their baseline variability. The modest correlation is perhaps not unexpected given that other factors, such as the gradient of a participant's cost landscape and their levels of sensory and motor noise, will affect the saliency of the cost landscape, and in turn the likelihood and degree of adaptation.

Experience with lower cost gaits initiates optimization

During the Second Experience Period, if just prior to the first probe participants were walking at low step frequencies, and thus experienced lower energetic costs, they appeared to initiate optimization and adapt toward the new optima (Fig. 4). However, if they were walking at high step frequencies, and thus experienced higher energetic costs, they rapidly returned to the initial preferred step frequency (Fig. 4). This difference in adaptation after low and high cost gaits was consistent regardless of whether the prior experience was self guided or metronome guided (P=0.03 for both). If instead participants were returned to the initial preferred step

Journal of Experimental Biology (2019) 222, jeb198234. doi:10.1242/jeb.198234

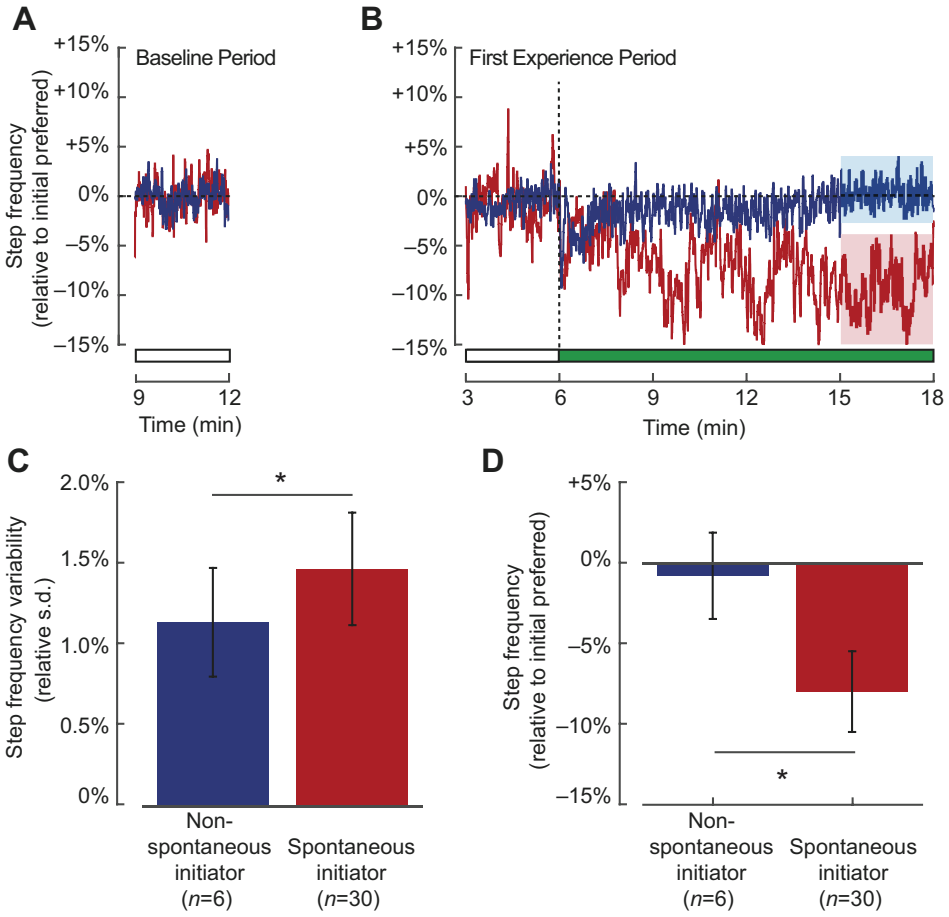

Fig. 3. Non-spontaneous and spontaneous initiators. (A) Self-selected step frequency during the final 3 min of the Baseline Period for a representative non-spontaneous initiator (blue) and spontaneous initiator (red). (B) Step frequency data during the First Experience Period for the same two representative participants. The horizontal bar indicates when the controller was turned on (green fill) and off (white fill). (C) Across all participants, spontaneous initiators displayed greater average step frequency variability than non-spontaneous initiators during the Baseline Period. (D) By the final 3 min of the First Experience Period, spontaneous initiators appeared to adapt their step frequency to converge on the energetic minima, while non-spontaneous initiators did not. Error bars represent 1 s.d. Asterisks indicate statistically significant differences between spontaneous and non-spontaneous initiators.

frequency immediately before the probe, as was the case with the metronome guided broad experience, they showed no adaptation (P=0.55; Fig. 4). This was despite these participants having broad experience with the cost landscape. It appears that providing participants with experience at a low cost gait and then allowing them to self-select their gait is sufficient for initiating optimization, while high cost experience and expansive experience with the new landscape is not. Importantly, the energy cost at the low cost gait is lower relative to the energy cost at the initially preferred step frequency under the new cost landscape, but not the original cost landscape (Fig. 1E). This indicates that the nervous system is updating its expectation of the energetic consequences of its gaits.

A local search strategy is used to converge on energetically optimal gaits

To investigate the interaction between high and low cost experience, as well as the order of the experience, we used the subset of participant data from the metronome guided discrete experience group that had experience with either the highest (+10%) or lowest (−15%) step frequency just prior to the first and last probes. Participants that had their first discrete experience at high step frequencies appeared to use prediction to rapidly move away from this high cost step frequency (Fig. 5A; Fig. S3A). They did so with an average time constant of 2.0±0.5 s, or about four steps. But their prediction was erroneous – having not yet experienced lower cost gaits, they returned to their initial preferred step frequency, which was suboptimal in the new cost landscape (1.7±2.3%, P=0.17; Fig. 5B; Fig. S3A). Participants that had their first discrete experience with low step frequencies more slowly descended the cost gradient, with an average time constant of 10.8±1.7 s, or about 20 steps (Fig. 5A). They eventually converged on the new optimal step frequency; the preferred step frequency during this probe of self-selected gait was significantly lower than the initial preferred step frequency (−6.6±1.5%, P=0.003; Fig. 5A; Fig. S3A) and not different from the optimal step frequency measured in our previous experiments (P=0.66; Selinger et al., 2015). This analysis is restricted to the first probe in participants whose first discrete experience was at −15% step frequency. This is a much lower frequency than the optimal frequency in the new cost landscape. Consequently, these participants had no prior explicit experience with the new optimum step frequency yet they converged to it (Fig. 5B). Gradual and sequential convergence to the new optima is consistent with a local search process. It is not consistent with a choose-best strategy, which would have required remaining at the −15% step frequency, or a sampling strategy, which would have required a broader sampling of a range of new gaits.

Optimization leads to new predictions of energy optimal gaits

During the last probe of self-selected gait, participants rapidly converged to the final preferred step frequency, regardless of the direction of prior experience (experience high: 2.8±0.5 s; experience low: 2.5±0.6 s; Fig. 5C,D; Fig. S3B). Also independent of the prior experience, the final preferred step frequency was not different from the optimal step frequency measured in our previous experiments (experience high: P=0.68; experience low: P=0.88). On average, participants' final preferred step frequency was −4.8±3.1%, significantly lower than the initial preferred step frequency (P=0.0015) and consistent with the expected optima. These results indicate that participants no longer

Journal of Experimental Biology (2019) 222, jeb198234. doi:10.1242/jeb.198234

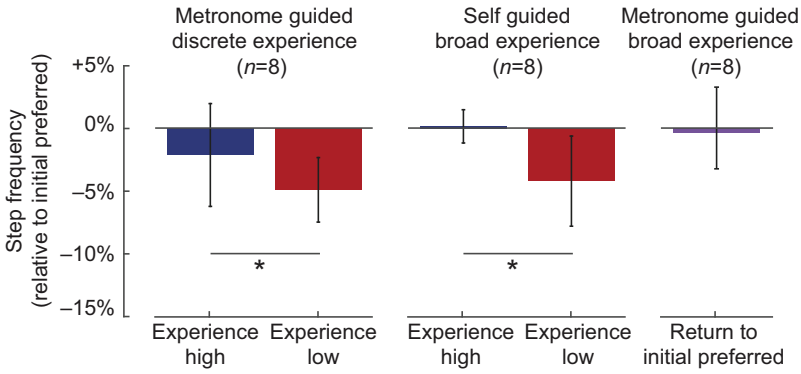

Fig. 4. Effect of experience direction on initiation of optimization. For each participant, we averaged data from the final 30 s of the first step frequency probe, and then averaged across participants. Error bars represent 1 s.d. Asterisks indicate statistically significant differences between experience high and experience low.

display slow adaptations consistent with optimization after experience with low cost gaits, but instead rapidly predict the optimal gait. They also indicate that participants' erroneous predictions after high cost gaits have been corrected and they now predict to the new cost optimum. The nervous system's optimization process appears to culminate in the formation of new predictions about optimal movements and the abolishment of old predictions. Consistent with this conclusion, when the controller was turned off and participants returned to the original cost landscape, they slowly unlearned the new prediction. With a time constant of 10.5± 1.8 s, they returned to a step frequency indistinguishable from their initial preferred step frequency (−0.8±3.0%, P=0.48).

Energy optimization as reinforcement learning

Our simple reinforcement learning model well describes the behaviour of our spontaneous initiators. We found that over about the same number of steps as our human participants, the model can converge on new energetically optimal gaits to achieve small cost savings (Fig. 6D). It also learns to predict the new cost landscape, rapidly returning to new cost optima when perturbed away, just as we found in our human experiments. When returned to the original and previously familiar cost landscape, it does not instantly remember old optima but instead has to unlearn its new prediction (Fig. 6E). This simple reinforcement learner cannot, however, explain the behaviour of our non-spontaneous initiators. Unlike the majority of our

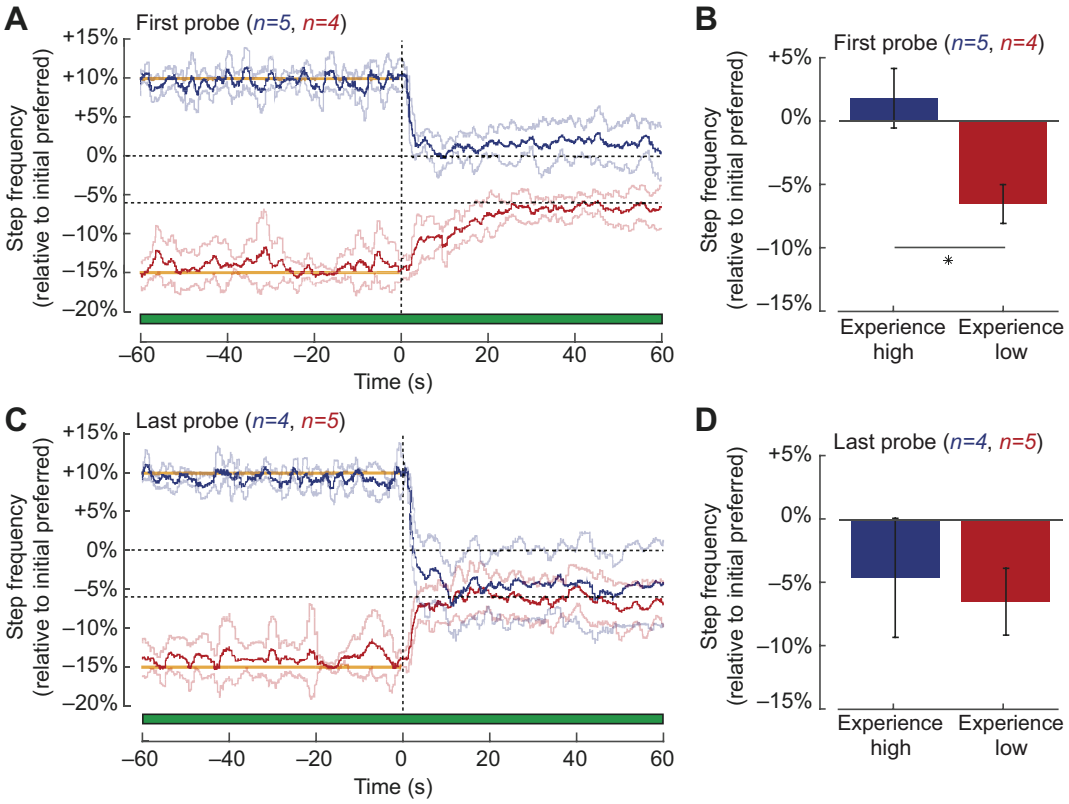

Fig. 5. Effect of experience direction during the first and last step frequency probe. (A,C) Step frequency time series data, averaged across participants, for the first (A) and last (C) probes following either experience high (blue) or experience low (red) step frequencies. The light blue and red lines represent 1 s.d. in step frequency for each time point. The horizontal bars indicate when the controller was turned on (green fill) and off (white fill), and the yellow lines indicate the prescribed metronome frequencies. (B,D) Steady-state step frequencies, averaged across participants, during the final 30 s of the probe for the first (B) and last (D) perturbations toward either high or low. Error bars represent 1 s.d. Asterisks indicate statistically significant differences between experience high and experience low.

Journal of Experimental Biology (2019) 222, jeb198234. doi:10.1242/jeb.198234

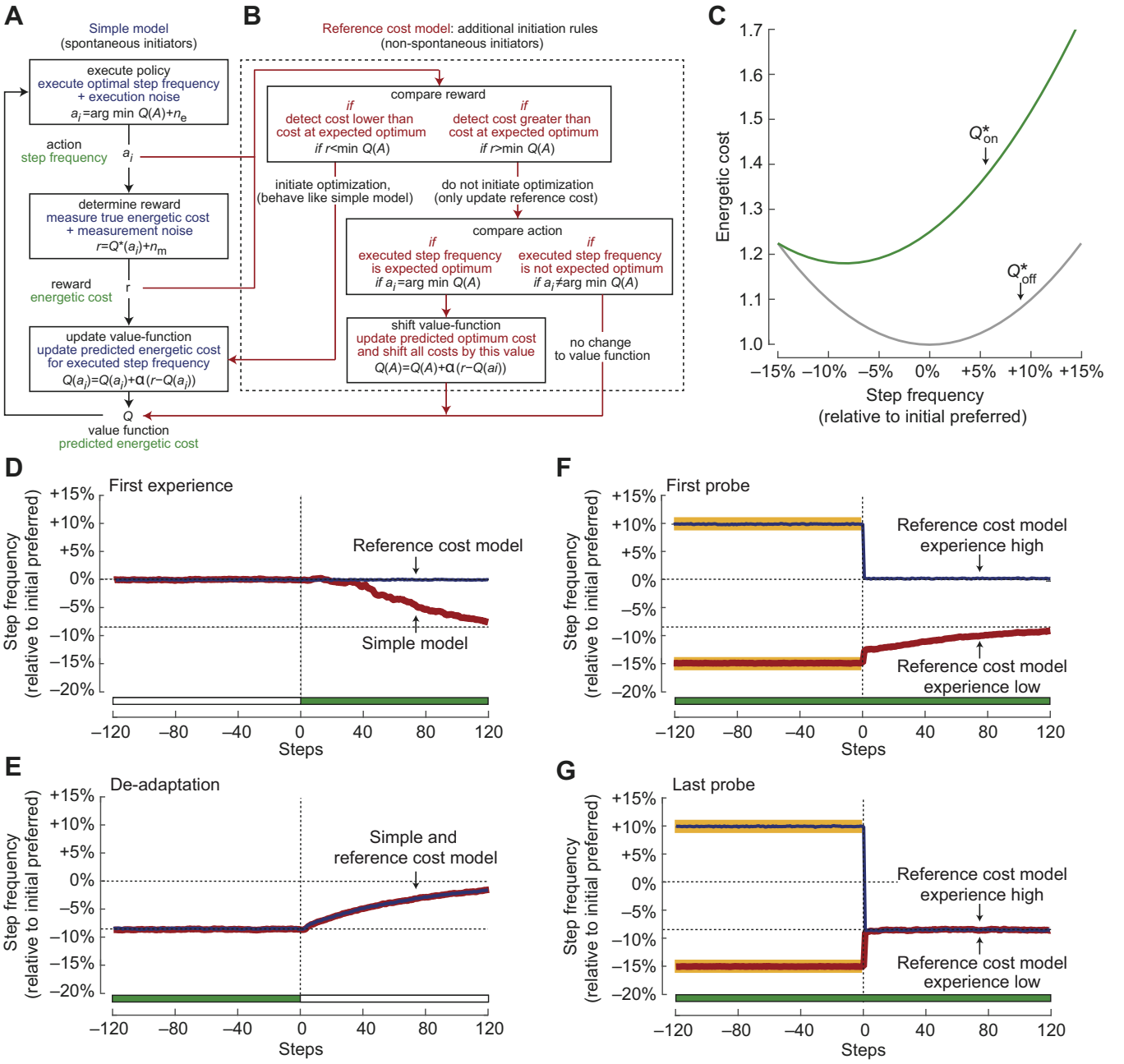

Fig. 6. Reinforcement learning model of energy optimization. (A) A simple model describing the behaviour of spontaneously initiating participants. (B) A more complex logic for updating the value function that prioritizes learning a reference cost and can describe the behaviour of non-spontaneously initiating participants. (C) The simulated energetic cost landscapes when the controller is turned off (Q^*_{off}) and on (Q^*_{on}). (D,E) Behaviour of the simple model of spontaneous initiators (red) and reference cost model of non-spontaneous initiators (blue) during the First Experience Period when the controller is first turned on (D) and the final de-adaptation period when the controller is turned off (E). (F,G) Behaviour of the reference cost model of non-spontaneous initiators during the first (F) and last (G) probes following experience with high (blue) and low (red) step frequencies. The horizontal bars indicate when the controller was turned on (green fill) and off (white fill), and the thick yellow lines indicate the prescribed frequencies prior to the probe.

experimental participants, this model will always spontaneously initiate optimization and begin converging on the optimal gait. This is true even for low learning rates where past predictions are much more heavily weighted over new measures (Fig. S2A).

The reference cost reinforcement learning model can capture many key behavioural features of our non-spontaneously initiating participants. First, it does not spontaneously initiate optimization (Fig. 6D). Second, it only initiates after experience in the new cost landscape with a frequency that has a lower cost than that at the initially preferred frequency. Third, after initiation, the algorithm gradually converges on the new optimum (Fig. 6F). Finally, much like our original model of spontaneous initiators, after convergence it can leverage prediction to rapidly return to the new optimum after a perturbation (Fig. 6G) but must slowly unlearn this optimum if returned to the original cost landscape (Fig. 6E).

DISCUSSION

Here, we used energy minimization in human walking to understand how the nervous system initiates and performs the optimization of its motor control strategies. We found that some participants tended

to explore, through naturally high gait variability, leading them to spontaneously initiate optimization. Others were more likely to exploit their current prediction of the optimal gait and required experience with lower cost gaits to initiate optimization. When optimization was initiated, participants gradually adapted their gait, in a manner consistent with a local search strategy, to converge on the new optima. Given more time and experience, this slow optimization was replaced by a new and rapid prediction of the optimal gait. These observed behaviours, where participants iteratively learn and then rapidly predict the new energy optimal gait, resemble the behaviours produced by our reinforcement learning algorithms. This suggests that the nervous system may use similar mechanisms to optimize gait for energy in walking, and perhaps optimize other movements for other objective functions.

Although our reinforcement learning models are simple, it is reasonable to ask whether even simpler algorithms could capture our experimental behaviour. One logical simplification would be to forgo the storing of the entire value function and only store the optimal gait and its associated cost. This simplified model would indeed capture our central behavioural observations by initiating optimization after experience with a lower cost, converging on a new energetic optimum using a local search, and rapidly predict the new optimum when perturbed away. Despite this, we prefer our slightly more complex models that learn value functions because we suspect they will better generalize to learning in the real world, for two reasons. First, storing information about non-optimal gaits seems valuable given that at times one may be constrained from using the globally optimal gait. For example, the no-value function model would need to relearn the optimal walking speed when constrained by a slow crowd that prevents walking at the globally optimal speed. In contrast, our value function models, which have memory of past non-optimal walking experience, would rapidly predict the new cost optimal speed in the face of this constraint (Pagliara et al., 2014; Snaterse et al., 2011). Ignoring this potentially useful past experience seems unlikely on the part of the nervous system, given that there will be times when it is energetically beneficial to recall it. Second, the simpler model avoids a value function only in the case where the learning task has one dimension, such as in our experimental paradigm. If instead, for example, the nervous system had to learn the optimal speed and step frequency, it would need to store the optimal step frequency, and its cost, at each speed (or vice versa). This is a one-dimensional value function for a two-dimensional optimization problem. As the nervous system cannot know *a priori* the dimensionality of the optimization problem, it may benefit from learning a high dimensional value function and then constraining the optimization problem depending on the constraints of the task.

Another logical simplification would be to forgo the updating of a reference cost prior to initiation of optimization. However, a model with this simplification does not reproduce key features of our experimental data. In our model of non-spontaneous initiators, prior to initiation of optimization, the learner only updates a reference cost (Fig. 6B). Without this feature, direct and gradual convergence to the new energetic optimum after forced experience with a low cost is not produced. Instead, because the reference cost has not been updated and therefore is expected to be that experienced under the controller off condition, this model will first rapidly shoot back to the old cost optimum after experience with a low cost. Only after updating this cost estimate, to its now higher cost value under the controller on condition, will it then gradually adapt to the new optimum. This updating of a reference cost prior to initiating optimization not only is necessary to reproduce our experimental findings but also has many parallels in neurophysiological habituation (Desmurget and Grafton, 2000; Shadmehr and Krakauer, 2008; Wolpert, 1997).

It is unclear from our experiments exactly what constitutes sufficient experience with a low cost gait to initiate optimization. For example, it may require a substantially lower cost, a sufficient number of steps at a lower cost, or some combination of these criteria. It is possible that high natural gait variability, as displayed by our spontaneous initiators, is in fact also triggering initiation through the updating of a reference cost because it provides sufficient experience with a low cost gait. If treated as so, all participants' behaviour could be explained by the reference cost model. However, deciphering an exact low cost experience criterion that fits all participants' behaviour is difficult, and perhaps not possible, as it probably varies across participants and is affected by additional factors such as the gradient of their cost landscape, their levels of sensory and motor noise, and their weighting of newly experienced costs. In addition, how the reward (or energetic cost) is measured (or sensed) is an open area of research (Dean, 2013; Wong et al., 2017). For the purposes of modelling, we assume that the criteria have been met during the experience with low cost prior to the first probe, in keeping with our experimental findings.

Principles of energetic optimality may determine the nervous system's balance between exploration and exploitation. Variability can aid with initiation by allowing the nervous system to locally sample a more expansive range of the cost landscape, clarify its estimate of the cost gradient, and identify the most promising dimensions along which to optimize (Herzfeld and Shadmehr, 2014; Tumer and Brainard, 2007; Wu et al., 2014). This variability may simply be a consequence of noisy sensorimotor control that fortuitously benefits the exploration process, or it may reflect intentional motor exploration by the nervous system (Tumer and Brainard, 2007; Wu et al., 2014). Recent work suggesting that humans actively reshape the structure of their motor output variability to elicit faster learning of reaching tasks is evidence of the latter (Wu et al., 2014). Learning rate also affects variability because new cost measurements are imperfect. The higher the learning rate, the greater the influence of the new and noisy cost measurements on the predicted optimal movement, resulting in more volatile predictions of the optimal gait and therefore more variable steps. This can accelerate learning of new optimal strategies in new contexts, reducing the penalty due to the accumulated cost of suboptimal movements during learning. But there is also a penalty to this high motor variability – once the new optimal strategy is learned, motor variability around this optimum means most movements are suboptimal. The optimal solution to this trade-off depends on how quickly the context is changing (Fig. 7). It is better to learn quickly and suffer steady-state variability about the new optimum when the context is rapidly changing. But, when the context changes infrequently, it is better to learn slowly and more effectively exploit the cost savings at the new optimum. Interestingly, the learning rate in our models, which we chose to match our experimental constraints, is optimal for a cost landscape that is changing approximately every 10 to 15 min, a rate of change not dissimilar from that applied in our experiment protocol. In humans, the nervous system probably has control over both the learning rate and the amount of exploration. It may adjust both based on its confidence in the constancy of the energetic conditions. This suggests that exploration, and potentially faster learning, could be promoted not through consistent experience in an energetic context but rather by experimentally alternating energetic contexts.

Journal of Experimental Biology (2019) 222, jeb198234. doi:10.1242/jeb.198234

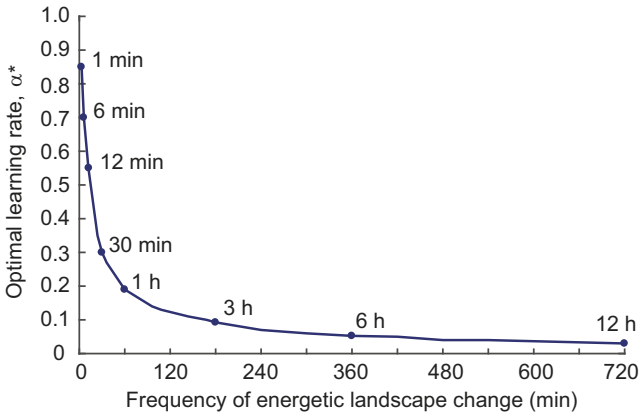

Fig. 7. Energetically optimal learning rates for varying frequency of cost landscape change. Measurement and execution noise were set at 2.0% and 1.5%, respectively.

Identifying the dimension of an optimization problem may be the trigger for initiation. The coordination of walking is a task of dauntingly high dimension (Bernstein, 1967; Scholz and Schöner, 1999). Various gait parameters, including walking speed, step frequency and step width must be selected, and numerous combinations of muscle activities can be used to satisfy any one desired gait. When presented with new contexts, the nervous system must identify which parameters, if any, to change in order to optimize its objective function. The difficulty of this task may partly explain why non-spontaneous adaptors do not initiate optimization when the exoskeletons are turned on and they are immediately shifted to a higher cost gait. Although it may be clear to the nervous system that energy costs are higher, it may remain unclear how it should change movement to lower the cost. This could also explain why in some past experiments, by our group (Selinger et al., 2015) and others (Reinkensmeyer et al., 2004; Zarrugh et al., 1974), participants did not initiate optimization and discover new energy optimal coordination strategies. Experience with lower step frequencies, and therefore lower costs, may have allowed the nervous system to identify that this is the relevant dimension along which to optimize. This behavioural phenomenon is captured by the addition of a reference cost to our simple reinforcement learning algorithm, and has parallels in classical feedback control models as well as neurophysiological habituation (Desmurget and Grafton, 2000; Shadmehr and Krakauer, 2008; Wolpert, 1997). Our experiments have demonstrated how the nervous system rapidly solves a one-dimensional optimization problem, where we alter the energetic consequences of a single gait parameter and apply targeted experience along this dimension of gait. How the identified mechanisms extend to optimizing higher dimension movement problems, like those often encountered in real-world conditions, remains an open area of research (Wong et al., 2019).

Prioritizing the learning of a reference cost, rather than constantly exploring new gaits, is perhaps a better general strategy for cost optimization in real-world conditions. Energetic cost continuously varies as conditions change in the real world, but unlike our experiment, only some conditions may benefit from the adoption of a new gait and exploring gaits away from the optimal gait comes with an energetic penalty. The continuous updating of a reference cost allows the nervous system to detect when there are reliable cost savings to be gained relative to the predicted optimal gait. It also allows the nervous system to compare differences between the two gaits and understand which walking adjustments led to the lower cost (Wolpert and Landy, 2012; Wolpert et al., 2011). This may allow the nervous system to learn the dimension along which exploration should proceed and quickly converge on the new optimal gait (Kording et al., 2007; Wolpert et al., 2011, 2001).

Unveiling the mechanisms that underlie the real-time learning of optimal movements may indicate how this process can be accelerated. This has direct applications in the development of rehabilitation programmes, the control of assistive robotic devices and the design of sport training regimes. A stroke patient faced with a change to their body, a soldier adapting to the new environment created by an exoskeleton and an athlete attempting to learn a novel task all seek new optimal coordination strategies. Our findings indicate that eliciting exploration through high motor variability as well as targeted experience along the relevant movement dimension could rapidly accelerate motor learning in these circumstances by cueing the nervous system to initiate optimization. Therapists and coaches may commonly be doing just this, based on years of accumulated knowledge about effective learning strategies. In this view, a more mechanistic understanding of the nervous system's internal algorithms could aid therapists and coaches in setting a course for a patient or athlete to navigate through various possible movement strategies.

Acknowledgements
We thank the SFU Locomotion Lab, T. J. Carroll and R. T. Roemmich for their helpful comments and suggestions.

Competing interests
The authors declare no competing or financial interests.

Author contributions
Conceptualization: J.C.S., J.D.W., S.N.S., J.M.D.; Methodology: J.C.S.; Formal analysis: J.C.S.; Investigation: J.C.S.; Writing - original draft: J.C.S.; Writing - review & editing: J.C.S., J.D.W., S.N.S., J.M.D.; Visualization: J.C.S.; Supervision: J.M.D.; Funding acquisition: J.M.D.

Funding
This work was supported by a Vanier Canadian Graduate Scholarship (J.C.S.), a Michael Smith Foundation for Health Research Fellowship (J.D.W.), the U.S. Army Research Office (grant W911NF-13-1-0268 to J.M.D.), and Natural Sciences and Engineering Research Council of Canada Discovery Grants (RGPIN-326825 to J.M.D. and RGPIN-2019-05677 to J.C.S.).

Supplementary information
Supplementary information available online at
http://jeb.biologists.org/lookup/doi/10.1242/jeb.198234.supplemental

References
Abram, S. J., Selinger, J. C. and Donelan, J. M. (2019). Energy optimization is a major objective in the real-time control of step width in human walking. *J. Biomech.* **91**, 85-91. doi:10.1016/j.jbiomech.2019.05.010

Alexander, R. M. (1996). *Optima for Animals*. Princeton University Press.

Atzler, E. and Herbst, R. (1928). Arbeitsphysiologische Studien III. *Pflugers Arch.* **215**, 292-328.

Bastian, A. J. (2008). Understanding sensorimotor adaptation and learning for rehabilitation. *Curr. Opin. Neurol.* **21**, 628-633.

Bellman, R. (1952). The theory of dynamic programming. *Proc. Natl Acad. Sci. USA* **38**, 716-719. doi:10.1073/pnas.38.8.716

Bernstein, N. A. (1967). *The Co-ordination and Regulation of Movements*. Pergamon Press.

Collins, S., Ruina, A., Tedrake, R. and Wisse, M. (2005). Efficient bipedal robots based on passive-dynamic walkers. *Science* **307**, 1082-1085. doi:10.1126/science.1107799

Dean, J. C. (2013). Proprioceptive feedback and preferred patterns of human movement. *Exerc. Sport Sci. Rev.* **41**, 36-43. doi:10.1097/JES.0b013e3182724bb0

Desmurget, M. and Grafton, S. (2000). Forward modeling allows feedback control for fast reaching movements. *Trends Cogn. Sci. (Regul. Ed.)* **4**, 423-431. doi:10.1016/S1364-6613(00)01537-0

Donelan, J. M., Kram, R. and Kuo, A. D. (2001). Mechanical and metabolic determinants of the preferred step width in human walking. *Proc. R. Soc. B* **268**, 1985-1992. doi:10.1098/rspb.2001.1761

Elftman, H. (1966). Biomechanics of muscle with particular application to studies of gait. *J. Bone Joint Surg. Am.* **48**, 363-377. doi:10.2106/00004623-196648020-00017
Flash, T. and Hogan, N. (1985). The coordination of arm movements: an experimentally confirmed mathematical model. *J. Neurosci.* **5**, 1688-1703. doi:10.1523/JNEUROSCI.05-07-01688.1985
Franklin, D. W. and Wolpert, D. M. (2011). Computational mechanisms of sensorimotor control. *Neuron* **72**, 425-442. doi:10.1016/j.neuron.2011.10.006
Herzfeld, D. J. and Shadmehr, R. (2014). Motor variability is not noise, but grist for the learning mill. *Nat. Neurosci.* **17**, 149-150. doi:10.1038/nn.3633
Kording, K. P., Tenenbaum, J. B. and Shadmehr, R. (2007). The dynamics of memory as a consequence of optimal adaptation to a changing body. *Nat. Neurosci.* **10**, 779-786. doi:10.1038/nn1901
Krakauer, J. W. (2006). Motor learning: its relevance to stroke recovery and neurorehabilitation. *Curr. Opin. Neurol.* **19**, 84-90. doi:10.1097/01.wco.0000200544.29915.cc
Krakauer, J. W. and Mazzoni, P. (2011). Human sensorimotor learning: adaptation, skill, and beyond. *Curr. Opin. Neurobiol.* **21**, 636-644. doi:10.1016/j.conb.2011.06.012
Kuo, A. D. and Donelan, J. M. (2010). Dynamic principles of gait and their clinical implications. *Phys. Ther.* **90**, 157-174. doi:10.2522/ptj.20090125
Lam, T., Anderschitz, M. and Dietz, V. (2006). Contribution of feedback and feedforward strategies to locomotor adaptations. *J. Neurophysiol.* **95**, 766-773. doi:10.1152/jn.00473.2005
Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D. and Wierstra, D. (2016). Continuous control with deep reinforcement learning. *arXiv preprint*. arXiv:1509.02971
Minetti, A. E., Ardigò, L. P. and Saibene, F. (1993). Mechanical determinants of gradient walking energetics in man. *J. Physiol.* **472**, 725-735. doi:10.1113/jphysiol.1993.sp019969
Molen, N. H., Rozendal, R. H. and Boon, W. (1972). Graphic representation of the relationship between oxygen-consumption and characteristics of normal gait of the human male. *Proc. K Ned. Akad. Wet. C* **75**, 305-314.
O'Connor, S. M. and Donelan, J. M. (2012). Fast visual prediction and slow optimization of preferred walking speed. *J. Neurophysiol.* **107**, 2549-2559. doi:10.1152/jn.00866.2011
Pagliara, R., Snaterse, M. and Donelan, J. M. (2014). Fast and slow processes underlie the selection of both step frequency and walking speed. *J. Exp. Biol.* **217**, 2939-2946. doi:10.1242/jeb.105270
Peters, J. and Schaal, S. (2008). Reinforcement learning of motor skills with policy gradients. *Neural Netw.* **21**, 682-697. doi:10.1016/j.neunet.2008.02.003
Ralston, H. J. (1958). Energy-speed relation and optimal speed during level walking. *Int. Z. Angew. Physiol.* **17**, 277-283. doi:10.1007/BF00698754
Reinkensmeyer, D., Aoyagi, D., Emken, J., Galvez, J., Ichinose, W., Kerdanyan, G., Nessler, J., Maneekobkunwong, S., Timoszyk, B., Vallance, K. et al. (2004). Robotic gait training: toward more natural movements and optimal training algorithms. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **7**, 4818-4821. doi:10.1109/IEMBS.2004.1404333
Scholz, J. P. Schöner, G. (1999). The uncontrolled manifold concept: identifying control variables for a functional task. *Exp. Brain Res.* **126**, 289-306. doi:10.1007/s002210050738
Schultz, W., Dayan, P. and Montague, P. (1997). A neural substrate of prediction and reward. *Science* **275**, 1593-1599. doi:10.1126/science.275.5306.1593
Scott, S. H. (2004). Optimal feedback control and the neural basis of volitional motor control. *Nat. Rev. Neurosci.* **5**, 532-545. doi:10.1038/nrn1427
Scott, S. H. and Norman, K. E. (2003). Computational approaches to motor control and their potential role for interpreting motor dysfunction. *Curr. Opin. Neurol.* **16**, 693-698. doi:10.1097/00019052-200312000-00008
Selinger, J. C., O'Connor, S. M., Wong, J. D. and Donelan, J. M. (2015). Humans can continuously optimize energetic cost during walking. *Curr. Biol.* **25**, 2452-2456. doi:10.1016/j.cub.2015.08.016
Shadmehr, R. and Krakauer, J. W. (2008). A computational neuroanatomy for motor control. *Exp. Brain Res.* **185**, 359-381. doi:10.1007/s00221-008-1280-5
Shadmehr, R., Huang, H. J. and Ahmed, A. A. (2016). A representation of effort in decision-making and motor control. *Curr. Biol.* **26**, 1929-1934. doi:10.1016/j.cub.2016.05.065
Simha, S. N., Wong, J. D., Selinger, J. C. and Donelan, J. M. (2019). A mechatronic system for studying energy optimization during walking. *IEEE Trans. Neural Syst. Rehabil. Eng.* **27**, 1416-1425. doi:10.1109/TNSRE.2019.2917424
Snaterse, M., Ton, R., Kuo, A. D. and Donelan, J. M. (2011). Distinct fast and slow processes contribute to the selection of preferred step frequency during human walking. *J. Appl. Physiol.* **110**, 1682-1690. doi:10.1152/japplphysiol.00536.2010
Srinivasan, M. and Ruina, A. (2005). Computer optimization of a minimal biped model discovers walking and running. *Nature* **439**, 72-75. doi:10.1038/nature04113
Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning*. MIT Press.
Sutton, R. S., Barto, A. G. and Williams, R. J. (1992). Reinforcement learning is direct adaptive optimal control. *IEEE Control Syst.* **12**, 19-22. doi:10.1109/37.126844
Todorov, E. (2004). Optimality principles in sensorimotor control. *Nat. Neurosci.* **7**, 907-915. doi:10.1038/nn1309
Todorov, E. and Jordan, M. I. (2002). Optimal feedback control as a theory of motor coordination. *Nat. Neurosci.* **5**, 1226-1235. doi:10.1038/nn963
Tumer, E. C. and Brainard, M. S. (2007). Performance variability enables adaptive plasticity of "crystallized" adult birdsong. *Nature* **450**, 1240-1244. doi:10.1038/nature06390
Umberger, B. R. and Martin, P. E. (2007). Mechanical power and efficiency of level walking with different stride rates. *J. Exp. Biol.* **210**, 3255-3265. doi:10.1242/jeb.000950
Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A. and Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *J. Exp. Psychol.* **143**, 2074-2081. doi:10.1037/a0038199
Wolpert, D. M. (1997). Computational approaches to motor control. *Trends Cogn. Sci. (Regul. Ed.)* **1**, 209-216. doi:10.1016/S1364-6613(97)01070-X
Wolpert, D. M. and Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nat. Neurosci.* **3**, 1212-1217. doi:10.1038/81497
Wolpert, D. M. and Landy, M. S. (2012). Motor control is decision-making. *Curr. Opin. Neurobiol.* **22**, 996-1003. doi:10.1016/j.conb.2012.05.003
Wolpert, D. M., Ghahramani, Z. and Flanagan, J. R. (2001). Perspectives and problems in motor learning. *Trends Cogn. Sci. (Regul. Ed.)* **5**, 487-494. doi:10.1016/S1364-6613(00)01773-3
Wolpert, D. M., Diedrichsen, J. and Flanagan, J. R. (2011). Principles of sensorimotor learning. *Nat. Rev. Neurosci.* **12**, 739-751. doi:10.1038/nrn3112
Wong, J. D., O'Connor, S. M., Selinger, J. C. and Donelan, J. M. (2017). Contribution of blood oxygen and carbon dioxide sensing to the energetic optimization of human walking. *J. Neurophysiol.* **118**, 1425-1433. doi:10.1152/jn.00195.2017
Wong, J. D., Selinger, J. C. and Donelan, J. M. (2019). Is natural variability in gait sufficient to initiate spontaneous energy optimization in human walking? *J. Neurophysiol.* **121**, 1848-1855. doi:10.1152/jn.00417.2018
Wu, H. G., Miyamoto, Y. R., Gonzalez Castro, L. N., Ölveczky, B. P. and Smith, M. A. (2014). Temporal structure of motor variability is dynamically regulated and predicts motor learning ability. *Nat. Neurosci.* **17**, 312-321. doi:10.1038/nn.3616
Zarrugh, M. Y., Todd, F. N. and Ralston, H. J. (1974). Optimization of energy expenditure during level walking. *Eur. J. Appl. Physiol. Occup. Physiol.* **33**, 293-306. doi:10.1007/BF00430237

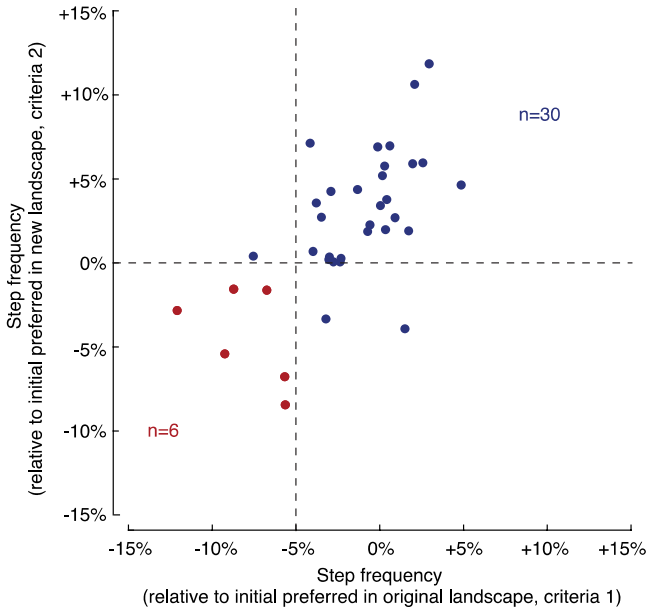

Fig. S1. Discrimination plot of spontaneous and non-spontaneous initiators. We defined spontaneous initiators as having a final step frequency during the First Experience Period consistent with the expected optima (-3SD from the initial preferred step frequency, or approximately -5%, x-axis), as well as displaying a significant change in step frequency from that displayed immediately after the exoskeleton was turned on (significantly different from 0%, y label). Although the above statistics, and not simple thresholds, were used for each criteria, the dashed lines illustrate roughly how each criteria divided the data.

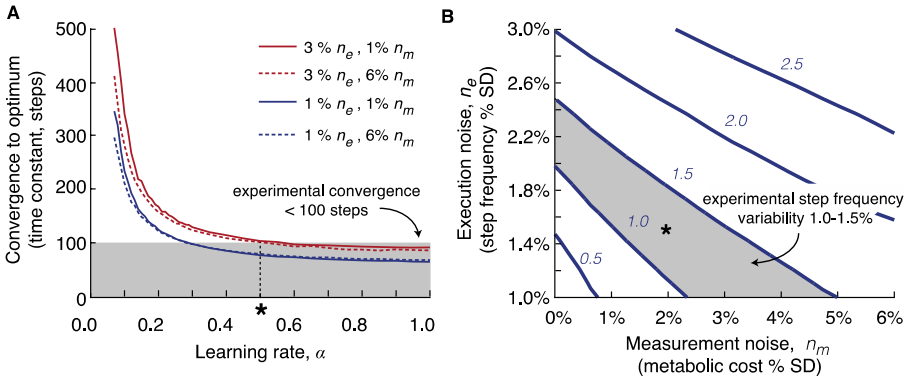

Fig. S2. Sensitivity analysis of model parameters. (A) Effect of varying the learning rate parameter on the rate of converge to the energetic optimum for different measurement and execution noise levels. The shaded region represents a reasonable convergence rate given that observed experimentally (maximum 100 steps), while the asterisk and dashed vertical line represents the chosen learning rate parameter value used in simulation (0.5). (B) Effect of varying measurement and execution noise on variability in steady state step frequency. Learning rate was kept constant at 0.5. Each line and the associated italic number represents a constant value of steady state step frequency. The shaded region represents reasonable steady state step frequencies given that observed experimentally (1.0% to 1.5%). The asterisk represents the chosen measurement and execution noise parameter values used in simulation (2.0% and 1.5%, respectively).

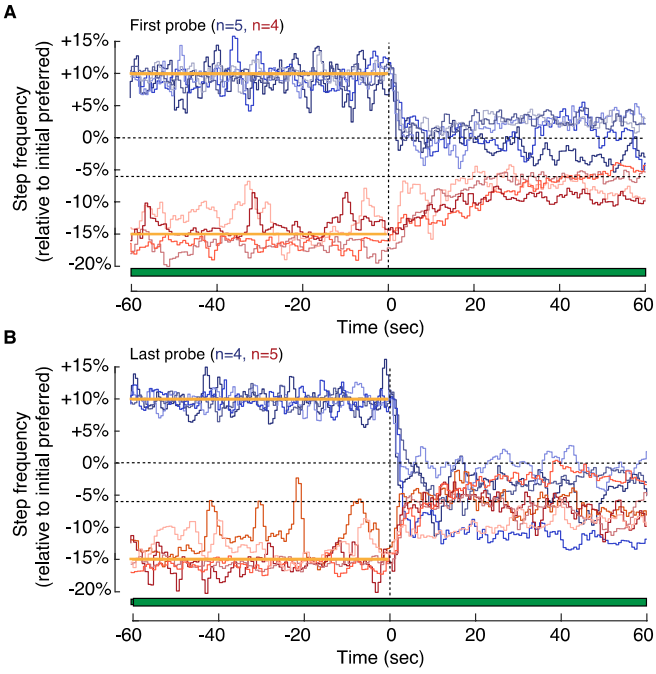

Fig. S3. Individual participant effect of experience direction during first and last step frequency probe. Step frequency time-series data for the first (A) and last (B) probes following experience either high (blue) or low (red) step frequencies for individual participants. The horizontal bars indicate when the controller is turned on (green fill) and off (white fill), and the yellow lines indicate the prescribed metronome frequencies.

Table S1. Participant numbers per protocol. We initially tested nine participants in each of the three Second Experience Periods. To account for a high number of spontaneous initiators in the Self Guided Broad Experience Period we added an additional two participants to this group to rebalance our conditions. To achieve the statistical power necessary to investigate the interaction between high and low cost experience, as well as the order of the experience, we added an additional seven participants to the Metronome Guided Discrete Experience group, one of which we found to be a non-spontaneous initiator. In total, we tested 36 participants, six of which were classified as spontaneous initiators and 30 which were non-spontaneous initiators.

Second Experience Period	Initial participants		Added to rebalance		Added to explore high low		Total participants		
	Spont.	Non-Spont.	Spont.	Non-Spont.	Spont.	Non-Spont.	Spont.	Non-Spont.	All
Metronome Guided Discreet	1	8	0	0	1	6	2	14	16
Metronome Guided Broad	1	8	0	0	0	0	1	8	9
Self Guided Broad	3	6	0	2	0	0	3	8	11
Total	5	22	0	2	1	6	6	30	36