Exploration-based learning of a step to step controller predicts locomotor adaptation

Nidhi Seethapathi[1,2,*], Barrett Clark[2], and Manoj Srinivasan[2,3,*]

[1]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA 19104, USA
[2]Mechanical and Aerospace Engineering, the Ohio State University, Columbus, OH 43210, USA
[3]Program in Biophysics, the Ohio State University, Columbus, OH 43210, USA
[*]snidhi@seas.upenn.edu, srinivasan.88@osu.edu

ABSTRACT

Humans are able to adapt their locomotion to a variety of novel circumstances, for instance, walking on diverse terrain and walking with new footwear. During locomotor adaptation, humans have been shown to exhibit stereotypical changes in their movement patterns. Here, we provide a theoretical account of such locomotor adaptation, positing that the nervous system prioritizes stability in the short timescale and improves energy expenditure over a longer timescale. The resulting mathematical model has two processes: a stabilizing controller which is gradually changed by a reinforcement learner that exploits local gradients to lower energy expenditure, estimating gradients indirectly via intentional exploratory noise. We consider this model walking and adapting under three novel circumstances: walking on a split-belt treadmill (walking with each foot on a different belt, each belt at different speeds), walking with an exoskeleton, and walking with an asymmetric leg mass. This model predicts the short and long timescale changes observed in walking symmetry on the split-belt treadmill and while walking with the asymmetric mass. The model exhibits energy reductions with exoskeletal assistance, as well as entrainment to time-periodic assistance. We show that such exploration-based learning is degraded in the presence of large sensorimotor noise, providing a potential account for some impairments in learning.

Introduction

Human locomotion combines versatility, stability, robustness to uncertainty, and energy economy in a manner not yet matched by analogous legged robots. Humans are able to adapt their gait to diverse terrain and locomotor circumstance, and the processes by which such adaptation proceeds are not completely understood, even at a behavioral and algorithmic level, let alone at the neural level[1,2]. Better understanding the principles governing locomotor adaptation may help us design better rehabilitation paradigms, build assistive or prosthetic robots that are more easily learnable or provide greater benefit[3–6], and more generally, accelerate motor learning[4,7].

Perhaps the most popular and well-studied experimental paradigm used to investigate human locomotor adaptation is walking on a ‘split-belt treadmill’[8–14]: a split-belt treadmill has two side-by-side belts that can be run at different speeds (Figure 1a). Walking on a split-belt treadmill, with one foot on each belt and with the two belts going at different speeds (Figure 1b), is a novel locomotor condition that humans have typically never experienced in their lives. Under this condition, humans initially walk quite asymmetrically, becoming more symmetric as time goes on (according to a particular symmetry metric). Numerous articles have recorded the myriad stereotypical gait adaptation phenomena during such walking. Specifically, when humans walk on a split-belt treadmill, they exhibit certain stereotypical changes in their left-right gait symmetry (Figure 1c-e), commonly characterized in the literature via the step length symmetry and the step time symmetry[8–11,15,16]. These symmetry changes are also accompanied by energy reductions over a longer period of time[12,14,17], suggesting that the final adaptation may be driven in part by energy minimization[1,18–20].

The locomotor adaptation proceeds over multiple time scales, short timescales of a few seconds (sometimes called early adaptation, seen as sharp jumps in symmetry in Figures 1c-e) to longer timescales of over tens of minutes (called late adaptation, seen as slow transients in Figures 1c-e). There may be other longer timescales of learning, but these have not been experimentally characterized). Here, we provide a first unified mathematical model to capture these human locomotor transients across timescales: the initial transient over the first few seconds, the longer transient over many tens of minutes, and finally the steady state adapted gait. This model takes the form of a local reinforcement learner that aims to walk stably while gradually reducing the energy cost of walking (Figure 2). We then examine other locomotor adaptation phenomena using the same framework, specifically, walking with a simplified exoskeleton and walking with an additional asymmetric leg mass.

Results

A minimal unified model of locomotor adaptation.

Energy economy and stability are two important criteria governing normal locomotion, whether on a treadmill or in the real world. Both for commonplace tasks[19–22] and for novel tasks[1,23,24], humans seem to move in a manner that minimizes energy consumption, perhaps for evolutionary reasons[25]. Similarly, walking stably, that is, without falling down is an important constraint as falls may result in injury. Normal walking is stabilized via feedback control, and we have previously characterized how foot placement and propulsion via the push-off impulse are modulated to produce a stable and robust walking motion[26–28]. These control actions respond to perturbations to the body (Figure 4), whether externally applied pushes, pulls, or trips[27,29,30] or internally generated sensorimotor noise[26,28], in a manner that a steady gait is restored over a few steps. We expect that short timescale responses to perturbations to walking are governed by this stabilizing controller, as without such fast timescale stabilization, the walker will fall. We expect this controller to remain unchanged as long as the treadmill belt speeds remain unchanged, or more generally, as long as the mechanical situation governing walking does not change. However, as soon as the belt speeds change, say to a split belt condition (unequal speeds), we posit that a higher level 'reinforcement learning' controller slowly changes the properties of the short-timescale controller to improve the energy economy of the gait by following an estimated gradient of the energy cost. Together, these two controllers may be visualized as an inner and an outer loop as shown in Figure 2, with the inner loop being the stabilizing controller and the outer loop being the learning controller or the reinforcement learner. We implement this two timescale controller in the context of a minimal biped model (Figure 4), starting with a stabilizing controller derived from human data[26–28]. See *Methods* for details of the biped model, the stabilizing controller, and the reinforcement learning controller.

Reinforcement learning simply means learning or performance improvement via intentional exploration of strategies through interactions with the environment and exploiting strategies that yield better performance, eventually resulting in an optimal control policy (or at least an improved control policy) for a given task, given some performance criterion. While there are numerous flavors of reinforcement learning[31–36], here, we use a simple flavor of reinforcement learning that is consistent with our understanding of locomotor control[2] (Figure 3). First, walking involves step to step variability. While some of this step to step variability is generated due to unavoidable sensory and motor noise, some of this variability is likely the nervous system's way of exploring the local neighborhood of control strategies – to enable learning[2,37]. By observing the energetic consequences of slightly different controller parameters on each step, the reinforcement learner slowly updates an estimate of the local gradient (derivatives or sensitivities) of the average energy cost with respect to the control actions. Our gradient estimator operates only on observed energy and body state, and does not have oracular access to the gradient directly nor does it do standard finite differencing (as such would not be physiological). This is 'local' or 'greedy' reinforcement learning, as the internal body parameters governing behavior are changed gradually, without large jumps to a different part of the parameter space. See the Methods section for more details regarding the gradient estimation and the reinforcement learning.

Split belt: Reinforcement learning of the stable biped captures the fast and slow symmetry transients

The aforementioned minimal biped model (Figure 4), when put through a split-belt adaptation protocol of Figure 1b — that is, equal belt speeds for a few minutes (baseline), then unequal speeds (adaptation phase), and then equal speeds again (deadaptation phase) — produces transients in walking asymmetry that are qualitatively identical to those found in experiment (Figure 5). Just like in the experiments, the step length asymmetry jumps to negative quickly at the beginning of adaptation and then slowly moves toward symmetry (zero) and then eventually positive asymmetry given enough time[14,17] or a fast enough learning rate. During de-adaptation, the step length asymmetry jumps to a more positive values and then slowly drifts back to zero. The step time asymmetry jumps to a positive value and remains positive during adaptation, and during deadaptation, jumps to a negative value and slowly drifts back to zero (symmetry).

Split belt: Initial fast timescale transients in symmetry are predicted by the stabilizing controller

The immediate fast timescale jumps in step length and step time asymmetry at the beginning of the adaptation and deadaptation phase (called early adaptation and early deadaptation, respectively) can be explained via the stabilizing controller. The foot placement controller of the biped directly adjusts the step length in response to fore-aft speed deviations, specifically reducing the step length if the body is vaulting over the foot too fast. Starting with the walking motion on the tied belt and suddenly changing the belt speeds to a split condition results in too fast a forward speed while on the slow belt and too slow a forward speed while on the fast belt (see *Supplementary Information*). These velocity deviations are transformed by the foot placement feedback controller into a smaller fast step length and a larger slow step length, resulting in the negative step length asymmetry.

The fast timescale response during deadaptation is similarly explained via how the forward speed on each belt compares to the nominal speed on that belt. An important difference is that at the end of the adaptation phase, the nominal forward speed on each belt is different, faster on the fast belt and slower on the slow belt. This means that when the belt speeds are tied again during the deadaptation phase, the foot placement controller reacts to the forward speed deviations from these asymmetric nominal forward speeds. The fast timescale transients in step time asymmetry are explained as a second order effect due to changes in the angle swept by each stance leg and the forward speeds on each belt.

Split belt: Steady state asymmetries are predicted by energy optimality

The slow transients due to the energy-reducing reinforcement learning control during the adaptation phase eventually approach positive step length asymmetry and retain the initial positive step time asymmetry (Figure 5). These asymmetries are also predicted by a separate optimization calculation that computes the two-step periodic walking motion that has the least total energy cost (see *Methods*). That is, this energy optimal two-step periodic motion has positive step length and step time asymmetry (Figure 6). Given that the reinforcement learning controller acts to reduce energy consumption, we find that the steady state of the reinforcement learning control agrees qualitatively with the pure periodic energy optimization. Because the gait is buffeted by internally generated noise during the reinforcement learning, the reinforcement learner implicitly performs a stochastic optimization. This explains the very small quantitative difference between the steady state of the reinforcement learning and a purely deterministic periodic energy optimum. Figure 6 also shows how the different components of the total metabolic cost contribute to the optimum: specifically, it is seen that the location of the optimum is determined by the dependence of the stance work on the gait asymmetry, specifically through the reduction of positive stance work by the leg on the fast belt[14,17]. The qualitative results (namely the signs of the optimal step length and step time symmetries) are robust to wide ranges of model parameters as well as leg swing or metabolic cost models, so as to be essentially independent of them. The generalized telescoping legged biped version of these steady state results, without restricting to the inverted pendulum limit, were presented in Seethapathi's thesis[18] and at Dynamic Walking (2018).

Split belt: Step time asymmetry converges to the steady state earlier than step length asymmetry.

During adaptation, step length asymmetry slowly goes from negative to positive, whereas the step time asymmetry remains positive after the initial fast timescale jump. Also, Figure 6 shows how the energy cost landscape has greater curvature along step time asymmetry rather than the step length asymmetry, again implying faster convergence for the energy optimizing gradient descent in the step time direction than in the step length direction. One recent study[38] with a shorter adaptation timescale observed that in that short timescale, split belt walking converged to the energy optimal step time but not the energy optimal step length. This may simply be because, as our model suggests, step time asymmetry may converge faster than step length asymmetry due to the aforementioned reasons.

Exoskeletal assistance: adaptation to state-dependent or time-periodic assistance.

There is an extensive literature on motor adaptation while wearing an assistive exoskeleton[3–6,39,40]. Here, we considered whether the learning process is able to adapt stably to such exoskeletal assistance. We assume a simplified exoskeleton[41] that torques the body forward about the ankle, which is equivalently to an ankle or hip exoskeleton for the minimal point-mass biped (Figure 7a).

We consider two types of exoskeletal assistance: (1) in which the exoskeleton produces forces based on the biped's body state and (2) in which the exoskeleton produces forces that are periodic in time. In the case of state-dependent assistance, which is the most common type of exoskeletal assistance[3–6], the exoskeleton applies a forward torque impulse when the leg reaches a particular angle. Figure 7b show that under this condition the learner adapts to a steady state that takes advantage of the provided assistance, resulting in a lower energy cost. In future work, we hope to examine how this adaptation is affected if the exoskeletal assistance is not precisely controlled (has step to step variability in either in magnitude or phase) or if the person and the exoskeleton are part of a human-in-the-loop optimization framework attempting to reduce the metabolic cost[4].

Exoskeletons providing time-periodic assistance are less common[39,40]. In the presence of such time-periodic perturbations and in the absence of learning (learning rate = 0), the system entrains to the period of the external assistance if the perturbation period is close enough the original gait period and if the perturbation is large enough, as in[39–41]. This entrainment is destroyed when learning is turned on with the learner not accounting for the perturbation. Entrainment to the perturbation is achieved when the internal model accounts for not just the body state but also the perturbation, for instance, including the perturbation timing as an input to the estimated state dynamics (see Methods section). This is because without it, an internal model that only considers body state would be inconsistent with the dynamics and can never accurately reflect the sensory data. Alternative to expanding the state of the internal model, using a slow enough learning rate or slow enough gradient updates (integrating over a sufficiently long past to estimate cost) also promotes entrainment. For instance, such low learning rates could be achieved if the learner uses a variable learning rate that turns down the learning rate when the internal model is not accurate.

Effect of adding asymmetric leg mass.

Prentice and Noble[42] showed how humans respond to an asymmetric mass added to just one of the two legs. Both step lengths and step times become immediately asymmetric and then gradually adapt toward symmetry[42,43]. And then, when the additional mass is removed, there in an after-effect in the opposite direction, which also decays slowly back to symmetry. Figure 8 shows an implementation how added mass affects the step length symmetry on our learning biped. Because the nominal biped model described above does not have explicit leg swing dynamics, the immediate response due to mass addition is due to enforcing a reduction in step length of the leg with the added mass, with the step length inversely proportional to the leg inertia. Such step length reduction is consistent with a leg swing controller that aims to produce a particular step length, but comes up short of its target on account of the altered mass (see appendix). After this immediate change in step length symmetry, the learner slowly changes the controller parameters in a manner that is consistent with experiment: that is, step lengths approach slowly converge toward symmetry.

Unresolved sensorimotor noise can degrade or destroy learning.

The learning described thus far relies on exploratory motor noise that perturbs the control policy on each step and builds a estimate of the gradient to improve the objective. This procedure relies on being able to estimate the state, the control, and the energy expenditure with reasonable accuracy. Increasing sensory noise in the measurements that go into the gradient estimate can degrade the accuracy of the gradient estimate, and large-enough sensory noise (comparable to the exploratory noise) can make the gradient estimate meaningless and destroy learning at any given learning rate. Unresolved motor noise has the same effect on learning. Say, we imagine that the motor noise consists of two components, one component that is ‘known’ to the nervous system (and can thus serve the purpose of exploratory noise) and another component (possibly partly due to the so-called signal dependent noise) that is unknown or not resolved by concurrent sensory information. When the latter component becomes too large, learning is degraded and eventually destroyed at any given learning rate. This effect of sensorimotor noise on learning may partly explain why learning may sometimes be impaired in populations with sensorimotor deficits[44–48].

Discussion

We have shown that the major phenomenology during adaptation and learning during a novel locomotor situation, namely walking on a split-belt treadmill, can be explained via a unified model that prioritizes stability in the short term and energy in the longer term.

While there are, by now, many tens of reinforcement learning-derived controllers for bipedal or quadrupedal simulations (whether robot-inspired or biologically plausible) and real physical robots (e.g.,[32–36]), no comparison has been previously made to formally compare the transients observed in a well-defined reinforcement learning process to that observed in human locomotor adaptation. More significantly, the flavors of reinforcement learning used in such machine learning literature are not directly relevant to the adaptation observed in continuous walking, as most of them usually involve or require episodic exploration: that is, walk for a while in the new environment, collect information, update parameters, and (needing to restart episodes of walking), rather than optimizing while continuously walking. Selinger and coworkers[2] outlined an elegant learning procedure for gradual continuous optimization of energy cost during walking via exploratory motor noise. This procedure was not designed to capture both short and long timescale transients, but only the long term transients, and did not rely on estimating and updating gradients. Further, in contrast to this procedure, we account for locomotor dynamics, specifically that the current energy cost not only depends on the action choices (step frequency in their case) but also the current transient state of the system. There have been a few other efforts to obtain gait adaptation in robots reminiscent of that in split belt walking, but via control mechanisms specifically engineered to obtain the observed adaptation, rather than naturally arise from other proto principles (e.g.,[49,50]).

Once the basic biped model and the learning algorithm was fixed, our theory had three free parameters, namely the learning rate, the internally generated exploratory noise, and internal model update rate (or memory used therein). It is well understood from optimization theory that too large learning rates will lead to instability of the descent procedure and too small learning rates may make no progress in the presence of noise. However, a range of intermediate learning rates will be adequate for effective locomotor learning. The specific learning rate chosen by the human will likely depend on the expected environmental changes[2], so that the learning rate may be higher when the nervous system believes that the environment may change more often or more rapidly. Similarly, the exploratory noise chosen by the nervous system for learning will depend on other sensorimotor or external noise that the nervous system cannot control or measure, so that the exploratory noise is large enough to obtain useful gradient information despite other noise. Indeed, it is known that motor variability may be helpful and be intentionally regulated to facilitate motor learning[2,37,51]. Thus, future experiments may test whether increasing frequency of environmental changes or other noise increases the learning rate and increases the exploratory noise (although it may be challenging to design an experiment to isolate just the exploratory noise).

The stabilizing feedback controller here had a feedforward component (nominal values for push off and step length) and a feedback component that operated on deviations from nominal or desired values of body state (e.g., forward velocity, forward position). Our primary results here rely on the reinforcement controller only operating on the feedforward component and the nominal values of the body state, and did not need to change the feedback gains to predict the phenomena. If these feedback gains are to be modified during learning, they need to be changed in a manner that safeguards the walker from going unstable. Allowing the feedback gains to change will allow the learning controller to converge to the stochastic optimum and may also allow for a longer lasting memory of the learned behavior.

We assume that during learning, the architecture of the stabilizing walking controller does not change, but only the parameters that define it. Thus, our learning is constrained by this assumed structure. Thus, we used the human derived controller as not just an initial condition for the control policy, but also an inductive bias for subsequent learning. Given that learning under such fixed structure is sufficient to explain many qualitative phenomena, we might hypothesize that these adaptation may not involve architectural changes to the neural control, but only small parametric changes (e.g., Hebbian learning).

In current and future work, we intend to describe how the model presented here generalizes to some other experiments involving locomotor adaptation: for instance, changing the energy landscape and gradient via speed or step frequency feedback (e.g.,[52–54]), split-belt with slopes[16], different deadaptation belt speeds[55], etc.), role of stored memory and experience in accelerating adaptation[2,52], and protocols to accelerate learning. We will also examine more complex high-DOF systems, learning under incomplete state information or delayed feedback, variable learning rates depending on internal model error, how the presence of many internal states in our learner promote 'savings' and anterograde interference[56–58], contrast error based learning versus minimizing an energy like objective, etc.

Methods

Minimal biped: dynamics, control, and objective

Dynamics. We consider a minimal model of a human (Figure 4a), consisting of a point-mass upper body and simple extensible legs that can apply forces on the upper body[59,60]. For this biped, the energy optimal walk on solid ground is the so-called inverted pendulum walking gait, in which the body vaults over the foot in a circular arc on each step (Figure 4b), with the transition from one step to the next achieved via push-off by the trailing leg, followed by a heel-strike at the leading leg (Figure 4c). The reinforcement learner uses this inverted pendulum gait, albeit on a split belt treadmill and this gait is entirely specified by the push-off impulses and the step length.

The total metabolic energy cost of walking for this biped is defined as a weighted sum of the positive and negative work done by the stance legs on the upper body and the work done to swing the legs forward[60,61]. This energy cost over a stride can then be normalized by the total time period of a stride to obtain the metabolic rate. The minimization of this metabolic rate is the objective of the reinforcement learner. This is a natural objective, as there is extensive evidence that human walk and run in a manner that approximately minimized energy consumption[1,20,21,24,59,60], even for short bouts and even under novel circumstances. See[59,60] for technical details; see[18] and the *Supplementary Information* on how to extend this model to walk on a split belt treadmill, specifically, defining leg work carefully. For simplicity and transparency, we choose to illustrate the technical issues in this well-studied irreducibly minimal low-dimensional model, whose global energy optimum in the absence of noise is easily obtained with great accuracy[59,60], rather than a more complex multibody multimuscle model of a human. The simplicity of the model allows us to test global convergence of learning in a manner that may not be possible with a much more complex biped.

The two key gait asymmetry metrics for this biped are step length asymmetry and step time asymmetry, defined as follows:

$$\text{Step length asymmetry} = \frac{D_{\text{fast}} - D_{\text{slow}}}{D_{\text{fast}} + D_{\text{slow}}} \text{ and Step time asymmetry} = \frac{T_{\text{fast}} - T_{\text{slow}}}{T_{\text{fast}} + T_{\text{slow}}} \tag{1}$$

where the fast and the slow step lengths (D_{fast} and D_{slow}) are defined at heel strike as in Figure 4c and the step times T_{fast} and T_{slow} are the respective stance times.

Stabilizing feedback control. As we are primarily interested in 2D phenomena here, we use the 2D version of the biped (but as an aside, we note that the model is generalizable to 3D walking). The biped has two control variables, namely, step length and push-off magnitude. These control variables are modulated to keep the biped stable, despite external or internal noisy perturbations and despite a change in the mechanical environment. e.g., walking on a split-belt treadmill instead of a regular treadmill. The values of these control variables on each step are decided by a discrete controller, as described below, derived from our prior human experiments[26–28]. Let us denote the two control variables together by the variable u. These have nominal values u_{nominal} that the biped uses in the absence of any perturbations. The body state at midstance is denoted by s and includes the forward position, the forward velocity and the running sum (i.e., discrete integral) of the forward position. These body states have nominal or 'desired' values s_{nominal} in the absence of any external perturbations, so the deviation from these nominal values are considered a perturbation to be corrected. The two control variables u_i at step i are changed by the following linear control rule as a function of the preceding midstance state s_i: $u_i = u_{\text{nominal}} + K \cdot (s_i - s_{\text{nominal}})$, where K is a matrix of feedback gains. The velocity dependence of the control gains ensures that the walker doesn't fall, the position dependence promotes station-keeping, and integral dependence ensures that the resulting system is robust to changes in the environment, namely, changing the belt speeds or going from a tied to a split treadmill. The three terms make the controller a discrete PID controller (proportional-integral-derivative). The default values for the control gain matrix K are obtained by fitting the dynamics of the model biped to the step to step map of normal human walking on a treadmill[26–28, 62].

Reinforcement learning to change controller parameters

We allow the parameters of the stabilizing controller to change slowly via a local reinforcement learning procedure, serving as an outer loop to the stabilizing controller as the inner loop (Figure 2). The parameters characterizing the stabilizing controller are u_{nominal}, K, and s_{nominal}. Here, we just allow u_{nominal} and the nominal forward velocity to change (part of s_{nominal}); without loss of generality due to translation invariance of the dynamics, we assume zero values for the nominal forward position and discrete sum of the forward position. We keep the feedback gains K fixed, not because we believe that they should remain constant, but because allowing them to change is not necessary to explain the central observed phenomena and is thus not parsimonious.

The reinforcement learning simulation is started off with the nominal default controller parameters corresponding to the tied belt condition. At each step, the reinforcement learner picks a new value of the controller parameters ($\bar{p}_{i+1}$, say), obtained as the sum of two terms: the old controller parameters from the previous step ($\bar{p}_i$) and a small step along the negative of the gradient estimate of the objective:

$$\bar{p}_{i+1} = \bar{p}_i - \alpha(g_i) + \nu_i, \tag{2}$$

where g_i is the current gradient estimate on the i^{th} step, the scalar α is a learning rate.

Rather than executing the next step using this new $\bar{p}_{i+1}$, we assume that the nervous system uses a perturbed version:

$$p_{i+1} = \bar{p}_{i+1} + \nu_i \tag{3}$$

where ν_i is the noise term, assumed to be uncorrelated Gaussian noise with some small standard deviation σ ($\nu_i \in \mathcal{N}(0, \sigma\mathbb{I})$). Here, we view the noise term ν_i as being exploratory, potentially 'intentionally' generated by the nervous system, allowing the nervous system to update the gradient. This noise term serves as 'persistent excitation' in the parlance of system identification[63]. In addition to this exploratory motor noise, there may be additional unavoidable noise that the nervous system cannot sense, which we ignore for now, but will add in later. Such additional unknown noise only serves to make the improvement in energy slower and more noisy, but do not change any of the qualitative results herein. Because the proposed reinforcement learning procedure directly operates on the parameters of the control policy, it is a type of policy gradient method (although we do not use the policy gradient theorem[64]), instead estimating the gradient as below entirely from exploratory steps. Because of the noise terms and because of the updating the gradient from limited data (see below), it is similar to a stochastic gradient descent on the control policy.

Gradient estimate and update

The human nervous system may not be able to directly estimate the gradient of reward with respect to actions, say, via a process analogous to automatic differentiation in machine learning. Instead, we expect the human nervous to build a model of the gradient via local exploration, updating a gradient estimate using only function evaluations. We find that the details of how this gradient update is achieved is not critical, only that the gradient estimate is a reasonable descent direction on average, that is, moving the control policy along the negative of the estimated gradient direction decreases the energy cost.

In the presented results, we use the following exploration-based gradient estimator, by simultaneously updating a model of the body's dynamics and the energy cost at each step. Informally, having a model of the body's dynamics allows the learner to estimate the longer-term consequences of the biped's actions. On step i, via some sensory estimation, the relevant body state (and possibly world states, if relevant) is estimated to be s_i and the metabolic rate over that step is estimated to be J_i. We posit that the nervous system maintains an internal dynamical model describing how these sensory estimates evolves, assumed linear for simplicity:

$$s_{i+1} = As_i + Bp_i + C, \quad \text{and} \quad J_{i+1} = Fs_i + Gp_i + H, \tag{4}$$

as a function of the control policy parameters p_i. These can be written in a single equation:

$$\begin{bmatrix} s_{i+1} \\ J_{i+1} \end{bmatrix} = \begin{bmatrix} A & B \\ F & G \end{bmatrix} \begin{bmatrix} s_i \\ p_i \end{bmatrix} + \begin{bmatrix} C \\ H \end{bmatrix}. \tag{5}$$

Upon estimating the current (s_i, J_i) on each step, the nervous system updates this linear internal model based in part on the error between the prediction from the internal model and the new measurements.

$$A \leftarrow A + \Delta A, \quad B \leftarrow B + \Delta B, \quad C \leftarrow C + \Delta C, \quad D \leftarrow D + \Delta D, \quad \ldots$$

This internal model update could be accomplished effectively many different ways[63]. For instance, they could be updated in a manner that latest data point is explained as well as possible by the updated model, and such an update is the so-called normalized least mean square filter. Another common way to update the linear model is the so-called recursive least squares, which essentially updates the model incrementally in a manner that the new model explains all the data ever seen or with a forgetting factor that weights new data more than old data. Here, we update the linear model in perhaps the simplest manner consistent with having finite memory: we update the linear model so as to best explains data over the previous few steps ($N = 30$, say). Some kind of finite memory, or forgetting of the original model or old data, is important so that the estimator may eventually converge to the new dynamic situation in a reasonable period.

This estimated internal model of state and energy cost dynamics contains information about the gradient of the metabolic rate with respect to the learning parameters p. Note that the matrix G is the gradient of the current metabolic rate with respect to the learning parameters. We hypothesize that the human prioritizes the long term or steady state metabolic rate J_∞. Given the internal model of the dynamics, the nervous system can estimate the consequences of parameter changes to the steady state (effectively simulating to steady state, say) and thus infer the relevant gradient. One can show that this gradient is: $\nabla J_\infty = G + F(I-A)^{-1}B$, where the first term gives the gradient of the short term energy cost over one step, while the second term corrects for the fact that the steady state s^* will be different from the current state s_i. Clearly, the nervous system could just as easily prioritize energy costs over an intermediate horizon by using the gradient of the mean energy cost over the next few steps to improve the control policy.

As noted, one can imagine a number of different ways for the nervous system to (effectively) estimate the relevant gradients, or indeed, to improve the control policy. The general qualitative phenomena explained in this manuscript is robust to many of these details. Therefore, we may not be able distinguish between the potentially many different ways for the nervous system may estimate and maintain gradient information. For instance, one variant of the gradient descent described in equation is the so-called gradient descent with 'momentum': $\bar{p}_{i+1} = \bar{p}_i - \alpha(q_i)$, where the descent direction q_i is not the current gradient g_i, but a running sum of current and past gradients as follows: $q_{i+1} = \beta q_i + (1-\beta)g_i$, with $0 < \beta \leq 1$. This procedure can inoculate the learning to noisy changes in the gradient, as well as speed up convergence to the optimum for appropriately chosen β[65].

Finally, our reinforcement learning method here also has a flavor of Q-learning, in that it tries to maintain a model of how both the current state and current action translate to future or long-term energy costs (equation 4). However, it is not precisely Q-learning due to the modeling choices made herein.

Steady state energy optimization

To study the steady state optimum separately, we perform a deterministic two-step optimization problem, computing the two-step periodic gait on the split-belt treadmill, determining the step length and push-off impulses on the two belts. The optimization methods are identical to those in our prior work[18,20,59,60]

References

1. Selinger, J. C., O Connor, S. M., Wong, J. D. & Donelan, J. M. Humans can continuously optimize energetic cost during walking. *Curr. Biol.* **25**, 2452–2456 (2015).

2. Selinger, J. C., Wong, J. D., Simha, S. N. & Donelan, J. M. How humans initiate energy optimization and converge on their optimal gaits. *J. Exp. Biol.* **222**, jeb198234 (2019).

3. Sawicki, G. S. & Ferris, D. P. Mechanics and energetics of level walking with powered ankle exoskeletons. *J. Exp. Biol.* **211**, 1402–1413 (2008).

4. Zhang, J. *et al.* Human-in-the-loop optimization of exoskeleton assistance during walking. *Science* **356**, 1280–1284 (2017).

5. Gordon, K. E. & Ferris, D. P. Learning to walk with a robotic ankle exoskeleton. *J. biomechanics* **40**, 2636–2644 (2007).

6. Cain, S. M., Gordon, K. E. & Ferris, D. P. Locomotor adaptation to a powered ankle-foot orthosis depends on control method. *J. neuroengineering rehabilitation* **4**, 1–13 (2007).

7. Emken, J. L. & Reinkensmeyer, D. J. Robot-enhanced motor learning: accelerating internal model formation during locomotion by transient dynamic amplification. *IEEE Transactions on Neural Syst. Rehabil. Eng.* **13**, 33–39 (2005).

8. Jensen, L., Prokop, T. & Dietz, V. Adaptational effects during human split-belt walking: influence of afferent input. *Exp. brain research* **118**, 126–130 (1998).

9. Reisman, D. S., Block, H. J. & Bastian, A. J. Interlimb coordination during locomotion: what can be adapted and stored? *J. neurophysiology* **94**, 2403–2415 (2005).

10. Choi, J. T., Vining, E. P., Reisman, D. S. & Bastian, A. J. Walking flexibility after hemispherectomy: split-belt treadmill adaptation and feedback control. *Brain* **132**, 722–733 (2008).

11. Bruijn, S. M., Van Impe, A., Duysens, J. & Swinnen, S. P. Split-belt walking: adaptation differences between young and older adults. *J. neurophysiology* **108**, 1149–1157 (2012).

12. Finley, J., Bastian, A. & Gottschall, J. Learning to be economical: the energy cost of walking tracks motor adaptation. *The J. physiology* **591**, 1081–1095 (2013).

13. Finley, J. M., Long, A., Bastian, A. J. & Torres-Oviedo, G. Spatial and temporal control contribute to step length asymmetry during split-belt adaptation and hemiparetic gait. *Neurorehabilitation neural repair* **29**, 786–795 (2015).

14. Sánchez, N., Simha, S. N., Donelan, J. M. & Finley, J. M. Taking advantage of external mechanical work to reduce metabolic cost: The mechanics and energetics of split-belt treadmill walking. *The J. physiology* **597**, 4053–4068 (2019).

15. Selgrade, B. P., Toney, M. E. & Chang, Y.-H. Two biomechanical strategies for locomotor adaptation to split-belt treadmill walking in subjects with and without transtibial amputation. *J. biomechanics* **53**, 136–143 (2017).

16. Sombric, C. J., Calvert, J. S. & Torres-Oviedo, G. Large propulsion demands increase locomotor adaptation at the expense of step length symmetry. *Front. physiology* **10**, 60 (2019).

17. Sanchez, N., Simha, S. N., Donelan, J. M. & Finley, J. M. Using asymmetry to your advantage: learning to acquire and accept external assistance during prolonged split-belt walking. *J. Neurophysiol.* (2020).

18. Seethapathi, N. *Transients, Variability, Stability and Energy in Human Locomotion*. Ph.D. thesis, The Ohio State University (2018).

19. Long, L. L. & Srinivasan, M. Walking, running, and resting under time, distance, and average speed constraints: optimality of walk–run–rest mixtures. *J. R. Soc. Interface* **10**, 20120980 (2013).

20. Seethapathi, N. & Srinivasan, M. The metabolic cost of changing walking speeds is significant, implies lower optimal speeds for shorter distances, and increases daily energy estimates. *Biol. letters* **11**, 20150486 (2015).

21. Ralston, H. J. Energy-speed relation and optimal speed during level walking. *Int Z angew Physiol einschl Arbeitsphysiol* **17**, 277–283 (1958).

22. Donelan, J. M., Kram, R. *et al.* Mechanical and metabolic determinants of the preferred step width in human walking. *Proc. Royal Soc. Lond. B: Biol. Sci.* **268**, 1985–1992 (2001).

23. Handford, M. L. & Srinivasan, M. Robotic lower limb prosthesis design through simultaneous computer optimizations of human and prosthesis costs. *Sci. reports* **6**, 19983 (2016).

24. Bertram, J. & Ruina, A. Multiple walking speed–frequency relations are predicted by constrained optimization. *J. theoretical Biol.* **209**, 445–453 (2001).

25. Srinivasan, M. Optimal speeds for walking and running, and walking on a moving walkway. *Chaos: An Interdiscip. J. Nonlinear Sci.* **19**, 026112 (2009).

26. Wang, Y. & Srinivasan, M. Stepping in the direction of the fall: the next foot placement can be predicted from current upper body state in steady-state walking. *Biol. Lett.* **10**, 20140405 (2014).

27. Joshi, V. & Srinivasan, M. A controller for walking derived from how humans recover from perturbations. *J. Royal Soc. Interface* **16**, 20190027 (2019).

28. Seethapathi, N. & Srinivasan, M. Step-to-step variations in human running reveal how humans run without falling. *ELife* **8**, e38371 (2019).

29. Hof, A., Vermerris, S. & Gjaltema, W. Balance responses to lateral perturbations in human treadmill walking. *J Exp Biol* **213**, 2655–2664 (2010).

30. Liu, C., Macedo, L. D. & Finley, J. M. Conservation of reactive stabilization strategies in the presence of step length asymmetries during walking. *Front. human neuroscience* **12**, 251 (2018).

31. Arulkumaran, K., Deisenroth, M. P., Brundage, M. & Bharath, A. A. Deep reinforcement learning: A brief survey. *IEEE Signal Process. Mag.* **34**, 26–38 (2017).

32. Haarnoja, T. *et al.* Learning to walk via deep reinforcement learning. *arXiv preprint arXiv:1812.11103* (2018).

33. Xie, Z., Berseth, G., Clary, P., Hurst, J. & van de Panne, M. Feedback control for cassie with deep reinforcement learning. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1241–1246 (IEEE, 2018).

34. Lee, D.-Y., Cho, Y.-H. & Lee, I.-K. Real-time optimal planning for redirected walking using deep q-learning. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 63–71 (IEEE, 2019).

35. Wang, J., Qin, W. & Sun, L. Terrain adaptive walking of biped neuromuscular virtual human using deep reinforcement learning. *IEEE Access* **7**, 92465–92475 (2019).

36. Peng, X. B., Berseth, G., Yin, K. & Van De Panne, M. Deeploco: Dynamic locomotion skills using hierarchical deep reinforcement learning. *ACM Transactions on Graph. (TOG)* **36**, 1–13 (2017).

37. Wu, H. G., Miyamoto, Y. R., Castro, L. N. G., Ölveczky, B. P. & Smith, M. A. Temporal structure of motor variability is dynamically regulated and predicts motor learning ability. *Nat. neuroscience* **17**, 312–321 (2014).

38. Stenum, J. & Choi, J. T. Step time asymmetry but not step length asymmetry is adapted to optimize energy cost of split-belt treadmill walking. *The J. Physiol.* **598**, 4063–4078 (2020).

39. Ahn, J. & Hogan, N. A simple state-determined model reproduces entrainment and phase-locking of human walking. *PloS one* **7**, e47963 (2012).

40. Ochoa, J., Sternad, D. & Hogan, N. Treadmill vs. overground walking: different response to physical interaction. *J. neurophysiology* **118**, 2089–2102 (2017).

41. Clark, B. C. *Energetic efficiency and stability in bipedal locomotion: 3D walking and energy-optimal perturbation rejection*. Ph.D. thesis, The Ohio State University (2018).

42. Noble, J. W. & Prentice, S. D. Adaptation to unilateral change in lower limb mechanical properties during human walking. *Exp. brain research* **169**, 482–495 (2006).

43. Hussain, S. J. & Morton, S. M. Perturbation schedule does not alter retention of a locomotor adaptation across days. *J. neurophysiology* **111**, 2414–2422 (2014).

44. Grau, J. W. *et al.* Learning to promote recovery after spinal cord injury. *Exp. neurology* 113334 (2020).

45. Welker, C. G., Voloshina, A. S., Chiu, V. L. & Collins, S. H. Shortcomings of human-in-the-loop optimization for an ankle-foot prosthesis: a case series. *bioRxiv* (2020).

46. Tyrell, C. M., Helm, E. & Reisman, D. S. Learning the spatial features of a locomotor task is slowed after stroke. *J. neurophysiology* **112**, 480–489 (2014).

47. Seuthe, J. *et al.* Split-belt treadmill walking in patients with parkinson?s disease: A systematic review. *Gait & posture* **69**, 187–194 (2019).

48. Hardwick, R. M., Rajan, V. A., Bastian, A. J., Krakauer, J. W. & Celnik, P. A. Motor learning in stroke: trained patients are not equal to untrained patients with less impairment. *Neurorehabilitation neural repair* **31**, 178–189 (2017).

49. Aoi, S. *et al.* Generation of adaptive splitbelt treadmill walking by a biped robot using nonlinear oscillators with phase resetting. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2274–2279 (IEEE, 2011).

50. Fujiki, S. *et al.* Adaptation mechanism of interlimb coordination in human split-belt treadmill walking through learning of foot contact timing: a robotics study. *J. Royal Soc. Interface* **12**, 20150542 (2015).

51. Tumer, E. C. & Brainard, M. S. Performance variability enables adaptive plasticity of ?crystallized?adult birdsong. *Nature* **450**, 1240–1244 (2007).

52. Snaterse, M., Ton, R., Kuo, A. D. & Donelan, J. M. Distinct fast and slow processes contribute to the selection of preferred step frequency during human walking. *J. Appl. Physiol.* **110**, 1682–1690 (2011).

53. Abram, S. J., Selinger, J. C. & Donelan, J. M. Energy optimization is a major objective in the real-time control of step width in human walking. *J. biomechanics* **91**, 85–91 (2019).

54. Simha, S. N., Wong, J. D., Selinger, J. C., Abram, S. J. & Donelan, J. M. Increasing the gradient of energetic cost does not initiate adaptation in human walking. *bioRxiv* (2020).

55. Vasudevan, E. V. & Bastian, A. J. Split-belt treadmill adaptation shows different functional networks for fast and slow human walking. *J. neurophysiology* **103**, 183–191 (2010).

56. Day, K. A., Leech, K. A., Roemmich, R. T. & Bastian, A. J. Accelerating locomotor savings in learning: compressing four training days to one. *J. neurophysiology* **119**, 2100–2113 (2018).

57. Roemmich, R. T. & Bastian, A. J. Two ways to save a newly learned motor pattern. *J. neurophysiology* **113**, 3519–3530 (2015).

58. Smith, M. A., Ghazizadeh, A. & Shadmehr, R. Interacting adaptive processes with different timescales underlie short-term motor learning. *PLoS Biol* **4**, e179 (2006).

59. Srinivasan, M. & Ruina, A. Computer optimization of a minimal biped model discovers walking and running. *Nature* **439**, 72–75 (2006).

60. Srinivasan, M. Fifteen observations on the structure of energy-minimizing gaits in many simple biped models. *J. R. Soc. Interface* **8**, 74–98 (2011).

61. Kuo, A. A simple model of bipedal walking predicts the preferred speed–step length relationship. *J. biomechanical engineering* **123**, 264–269 (2001).

62. Perry, J. A. & Srinivasan, M. Walking with wider steps changes foot placement control, increases kinematic variability and does not improve linear stability. *Royal Soc. open science* **4**, 160627 (2017).

63. Goodwin, G. C. & Sin, K. S. *Adaptive filtering prediction and control* (Courier Corporation, 2014).

64. Sutton, R. S., McAllester, D. A., Singh, S. P. & Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, 1057–1063 (2000).

65. Goodfellow, I., Bengio, Y., Courville, A. & Bengio, Y. *Deep learning*, vol. 1 (MIT press Cambridge, 2016).

Acknowledgements

NS was funded by a Schlumberger fellowship grant and Ohio State University travel and research awards. MS was supported by NSF CMMI grant no. 1254842.

Author contributions statement

N.S. and M.S conceived the study, created the mathematical models, performed the computations, analyzed the results, and wrote the manuscript. All authors reviewed the manuscript. B.C. contributed to adaptation to state-dependent and time-periodic exoskeletal assistance.

Additional information

Competing interests. The authors declare that they have no competing interests.

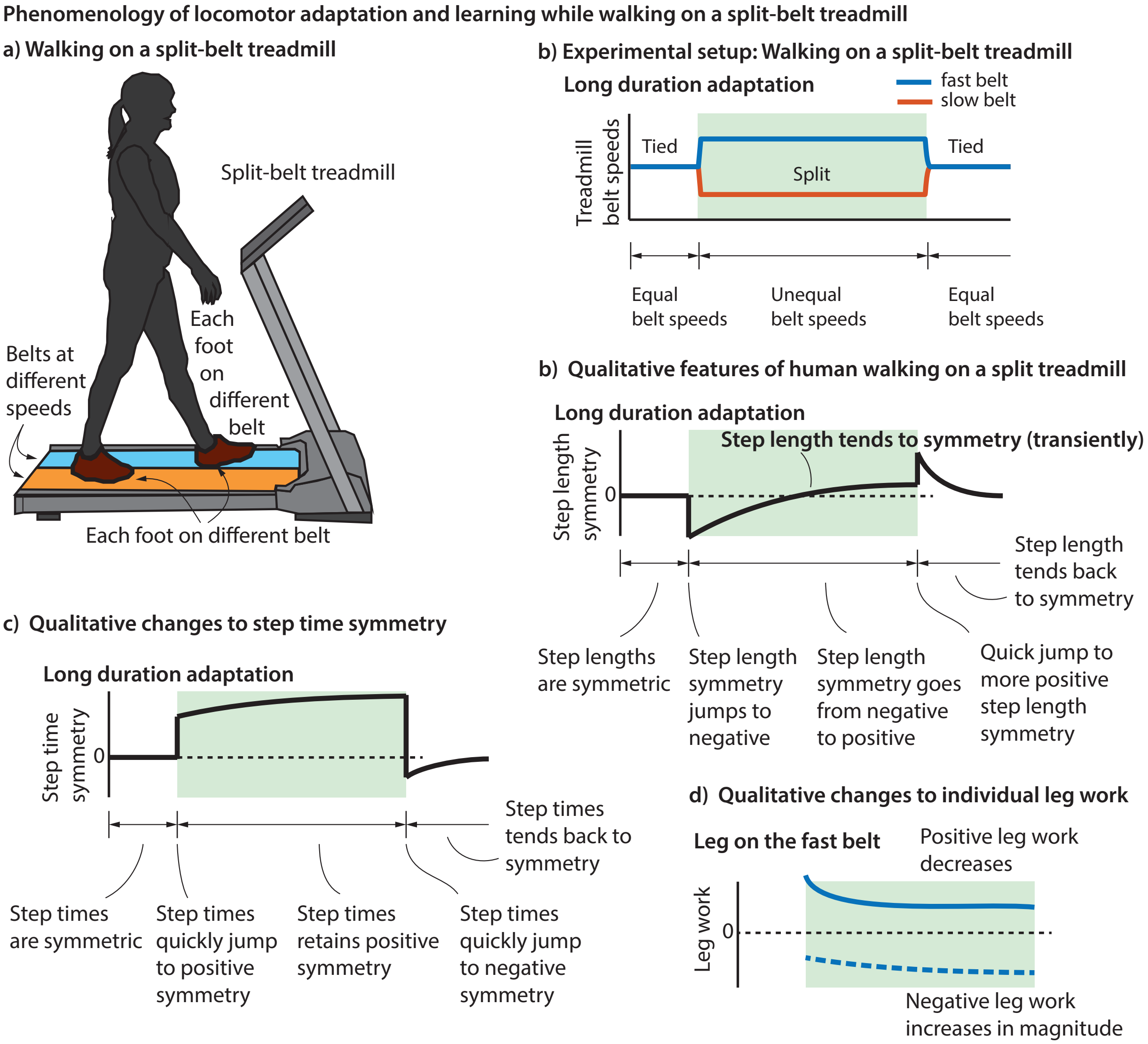
Phenomenology of locomotor adaptation and learning while walking on a split-belt treadmill
a) Walking on a split-belt treadmill
Split-belt treadmill
Each foot on different belt
Belts at different speeds
Each foot on different belt
b) Experimental setup: Walking on a split-belt treadmill
Long duration adaptation
fast belt
slow belt
Treadmill belt speeds
Tied
Split
Tied
Equal belt speeds
Unequal belt speeds
Equal belt speeds
b) Qualitative features of human walking on a split treadmill
Long duration adaptation
Step length tends to symmetry (transiently)
Step length symmetry
0
Step length tends back to symmetry
Step lengths are symmetric
Step length symmetry jumps to negative
Step length symmetry goes from negative to positive
Quick jump to more positive step length symmetry
c) Qualitative changes to step time symmetry
Long duration adaptation
Step time symmetry
0
Step times tends back to symmetry
Step times are symmetric
Step times quickly jump to positive symmetry
Step times retains positive symmetry
Step times quickly jump to negative symmetry
d) Qualitative changes to individual leg work
Leg on the fast belt
Positive leg work decreases
Leg work
0
Negative leg work increases in magnitude

Figure 1. Walking in an unusual asymmetric environment. a) Walking on a split-belt treadmill is a standard paradigm used to study how humans adapt and learn to walk in a novel environment. The two feet are on separate belts, which can be at equal or unequal speeds. b) A typical split-belt adaptation protocol involves starting with equal belt speeds (tied belts), then unequal belt speeds (split-belt) and then equal speeds again. c) Step lengths are approximately symmetric in the tied condition, jumps quickly to negative step length asymmetry as soon as the belt speeds become unequal. Then the step length asymmetry slowly rises toward symmetry and given sufficient time for adaptation, eventually reaches positive step length asymmetry. When the belts are tied again, the step length asymmetry jumps quickly to an even more positive value and then slowly approaches symmetry. d) Step time asymmetry jumps from symmetry to positive asymmetry at the beginning of the adaptation phase, remains positive through the phase. When the treadmill is tied again, the step time asymmetry quickly jumps to negative before eventually tending to symmetry. e) During split-belt adaptation, the positive leg work by the fast leg reduces, whereas the negative leg work may increase in magnitude. Author's note: In the next version of this preprint, we will overlay these idealized qualitative trends in the figure with experimental data digitized from prior studies.

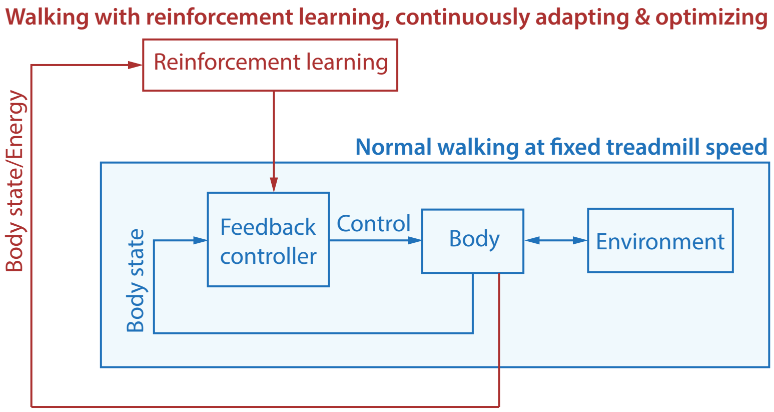
Walking with reinforcement learning, continuously adapting & optimizing
Reinforcement learning
Body state/Energy
Normal walking at fixed treadmill speed
Feedback controller
Control
Body
Environment
Body state

Figure 2. Human adaptation as stabilizing control and reinforcement learning. We propose that human response to a novel environment can be conceptually divided into two aspects: (1) the inner loop (blue), representing a fast timescale response due to the stabilizing controller, aimed at preventing falling and station keeping; (2) the outer loop (red), representing slower learning-like processes that tune parameters of the inner loop controller to lower the energy cost of walking (or possibly another performance metric).

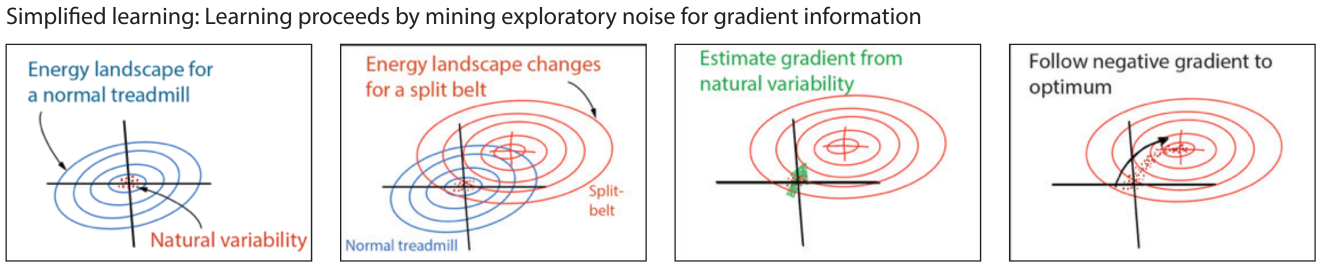
Simplified learning: Learning proceeds by mining exploratory noise for gradient information
Energy landscape for a normal treadmill
Natural variability
Energy landscape changes for a split belt
Split-belt
Normal treadmill
Estimate gradient from natural variability
Follow negative gradient to optimum

Figure 3. Reinforcement learning via exploratory noise for gradient estimation. Exploratory noise in the control policy parameters (partly accounting for natural variability) allows the learner to estimate the gradient of the objective with respect to the control parameters[2]. Then, the learner can follow the negative of this gradient to reduce the value of the objective.

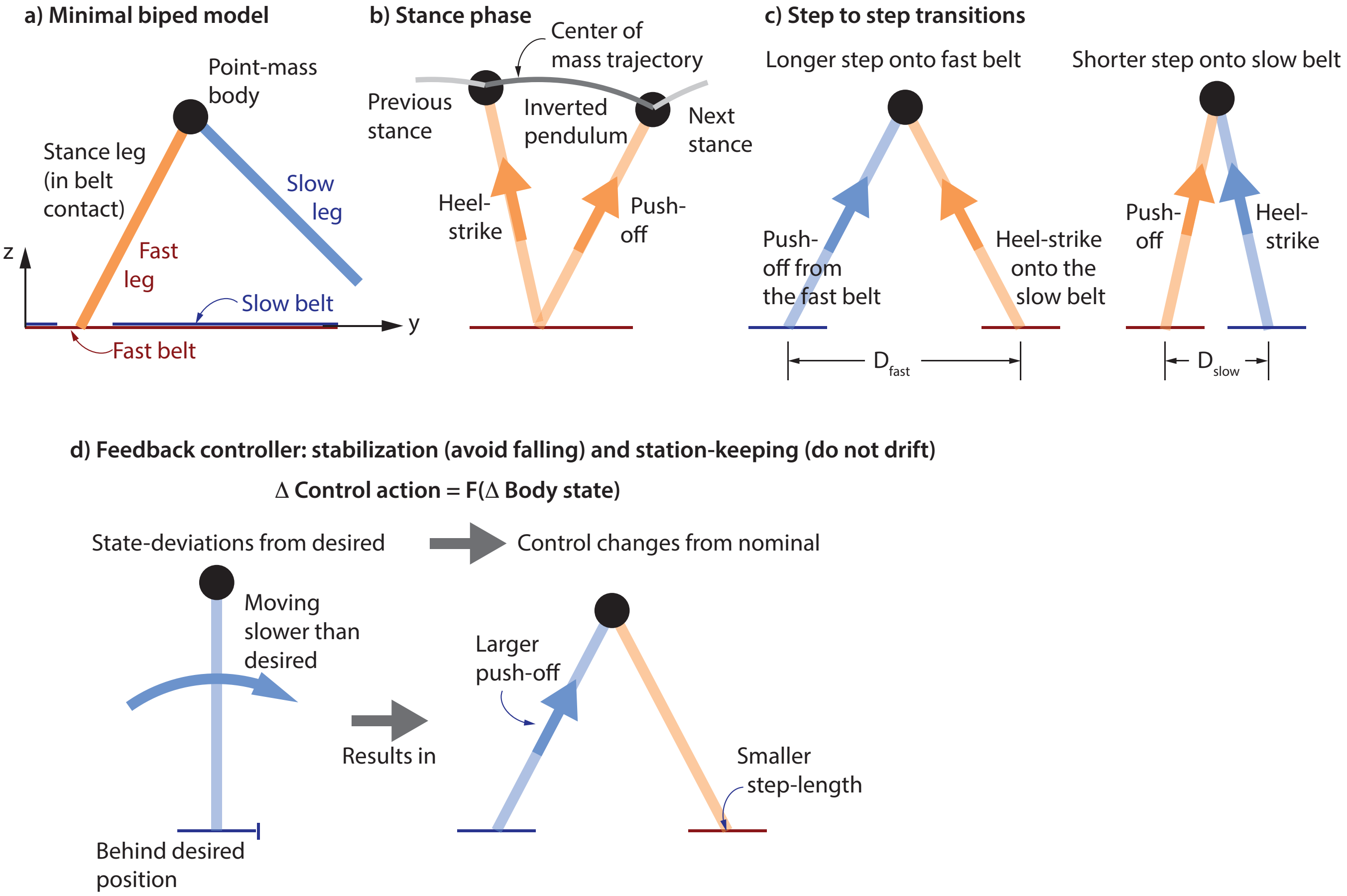
a) Minimal biped model
Point-mass body
Stance leg (in belt contact)
Slow leg
Fast leg
z
y
Slow belt
Fast belt
b) Stance phase
Center of mass trajectory
Previous stance
Inverted pendulum
Next stance
Heel-strike
Push-off
c) Step to step transitions
Longer step onto fast belt
Shorter step onto slow belt
Push-off from the fast belt
Heel-strike onto the slow belt
Push-off
Heel-strike
D_{fast}
D_{slow}
d) Feedback controller: stabilization (avoid falling) and station-keeping (do not drift)
Δ Control action = F(Δ Body state)
State-deviations from desired
Control changes from nominal
Moving slower than desired
Behind desired position
Results in
Larger push-off
Smaller step-length

Figure 4. Minimal walking model. a) A minimal biped model with a point-mass body performing an inverted pendulum walking gait. b) During single stance phase, when only one leg contacts the belt, the body vaults over the foot like an inverted pendulum. c) Step to step transitions from one inverted pendular stance phase to the next are accomplished via a push-off and heel-strike impulse. d) The biped is prevented from falling backward or drifting off the treadmill by a stabilizing feedback controller that modify the push-off impulse and the step length in response to perturbations: for instance, using a smaller step length and a larger push-off when the forward velocity is too low or if the forward position trails the preferred position.

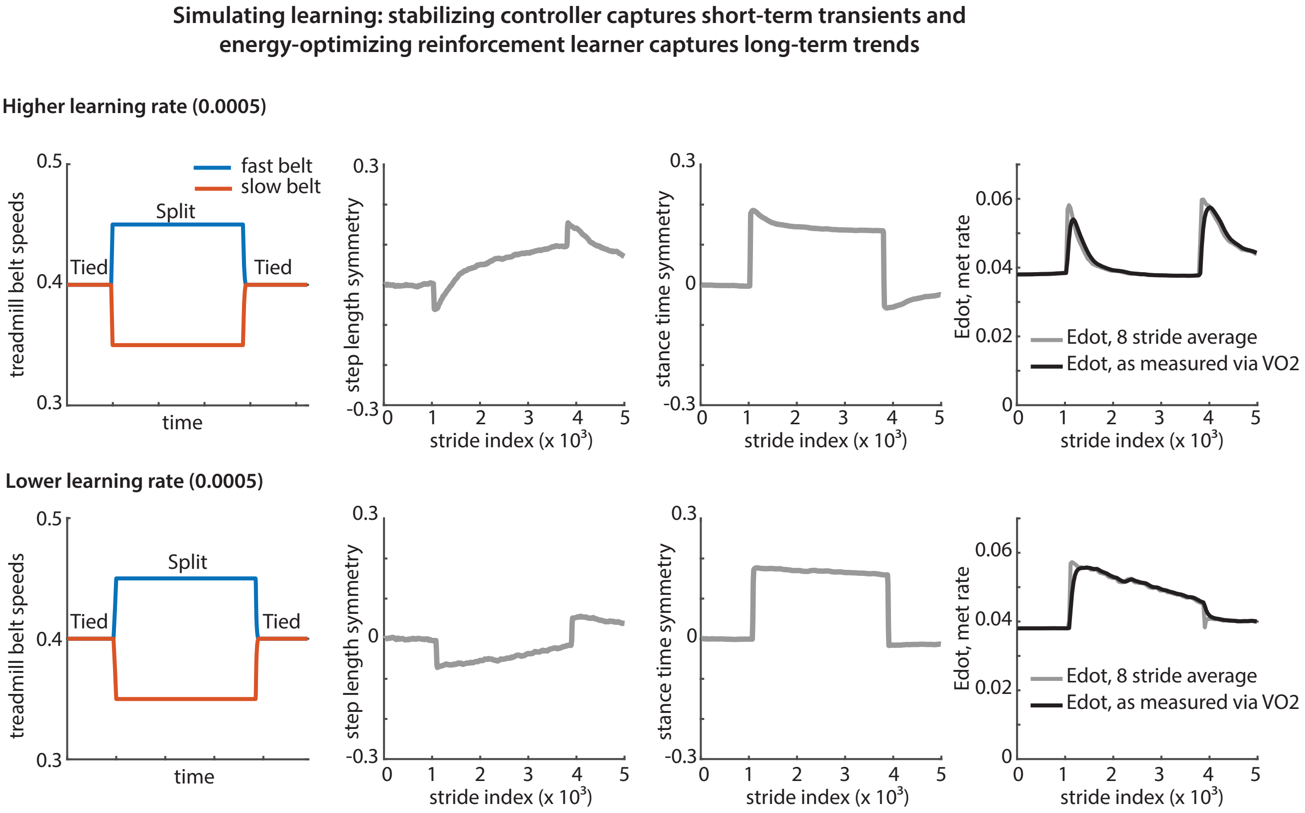

Figure 5. Policy gradient reinforcement learning predicts observed slow and fast transients, and steady state. We see that the step length asymmetry and the step time asymmetry show the same trends as in the experiment. The step time asymmetry jumps to positive and remains positive during adaptation, whereas it jumps to negative and slowly trends to zero during de-adaptation. The step length asymmetry jumps to negative and slowly increases to positive (a) high learning rate) or zero (b) low learning rate) during adaptation, and jumps to a more positive value and then trends to zero during de-adaptation. Author's note: In the next version of this preprint, we will show results from an intermediate learning rate that is chosen so as to fit the observed locomotor transient timescales. We will also add labels or experimental data to this figure to show how these trends are analogous to those in Figure 1.

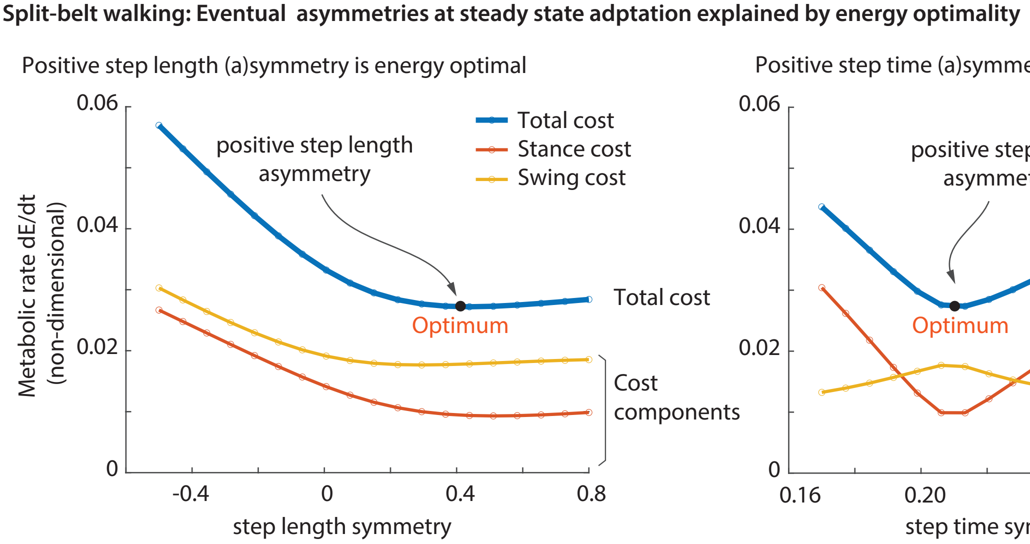

Figure 6. Optimal steady state split-belt walking explained by energy optimality. a) Positive step length asymmetry is energy optimal. b) Positive step time asymmetry is energy optimal. The various components of the total model metabolic cost are shown, namely, the cost of mechanical work performed by the legs during stance and the total swing costs. Cost due to individual leg work, showing that the fast leg work is lower and slow leg work is higher, was shown in Seethapathi's thesis[18] with a telescoping legged biped, without restricting to the inverted pendulum limit.

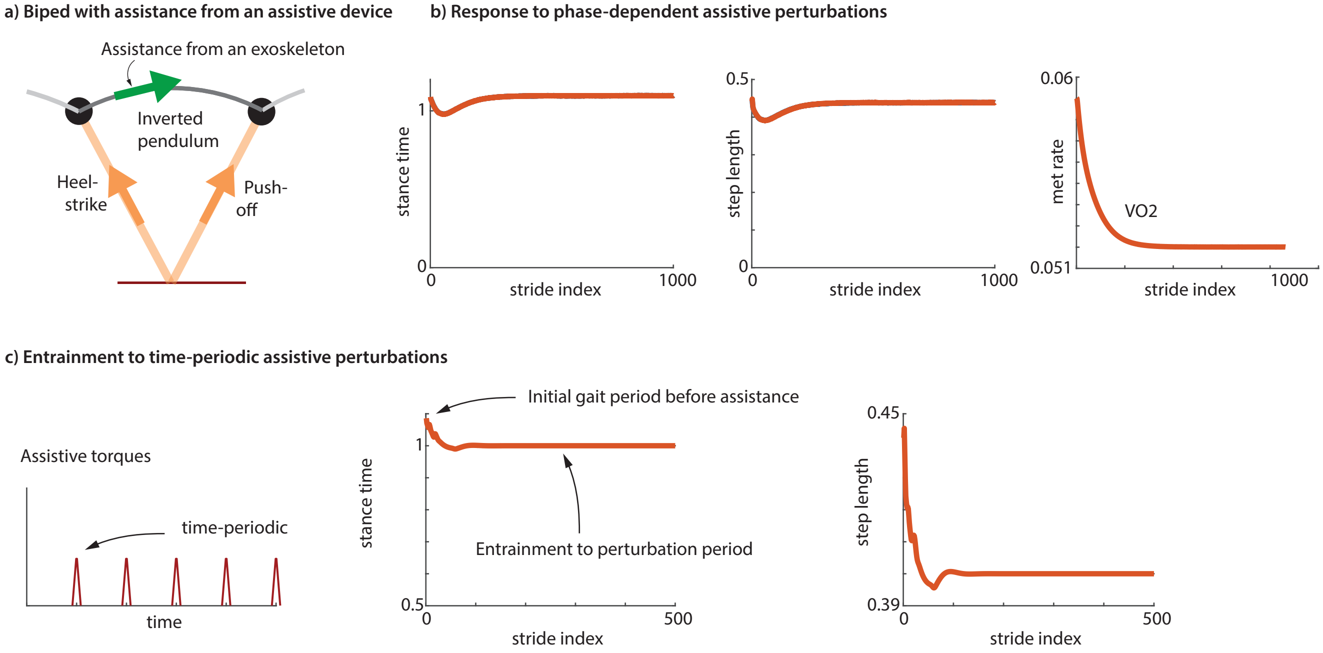

Figure 7. Adapting to exoskeleton assistance. a) A simple exoskeleton that provides forward propulsive torque. b) Learning biped adapts stably to the assistive torques being applied at a fixed phase during each step. The step time and step length remain symmetric throughout and approach new steady states. c) Learning biped entrains to time-periodic perturbations when it pays attention to the timing of the periodic perturbations.

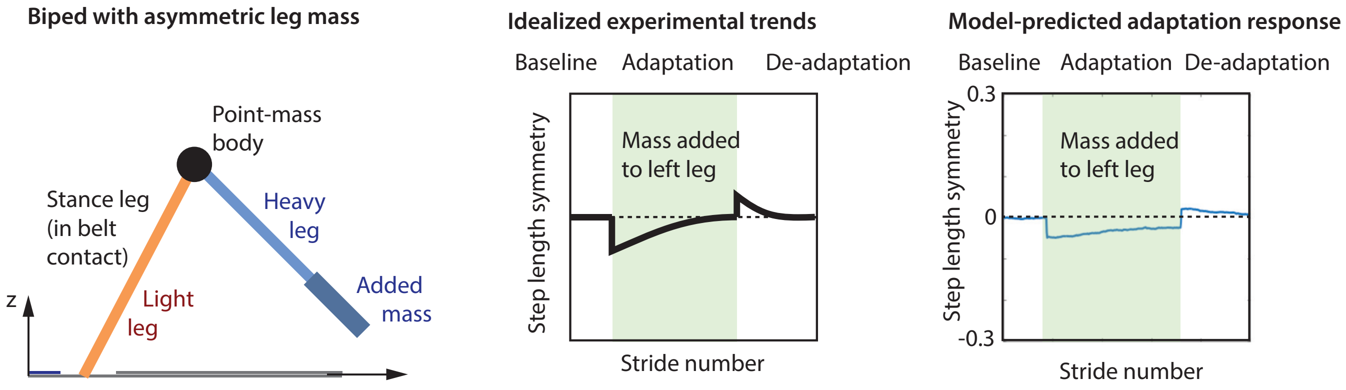

Figure 8. Adaptation to asymmetric leg mass. Adding an asymmetric mass to just one of the two legs results in stereotypical adaptation and deadaptation in step length symmetry that is recapitulated in the biped model.