Received: 23 June 2020 | Revised: 25 September 2020 | Accepted: 29 September 2020

DOI: 10.1002/bies.202000157

BioEssays WILEY

PROBLEMS & PARADIGMS

Prospects & Overviews

Taming fitness: Organism-environment interdependencies preclude long-term fitness forecasting

Guilhem Doulcier[1] | Peter Takacs[1] | Pierrick Bourrat[2,1]

[1] Department of Philosophy & Charles Perkins Centre, The University of Sydney, Sydney, New South Wales 2006, Australia

[2] Department of Philosophy, Macquarie University, Sydney, New South Wales 2109, Australia

Correspondence
Pierrick Bourrat, Department of Philosophy, Macquarie University, Sydney, New South Wales, Australia.
Email: p.bourrat@gmail.com

Funding information
GD's and PB's research was supported by a Large Grant from the John Templeton Foundation (Grant ID 60811). PB's research was also supported by a Macquarie University Reseearch Fellowship. PT's research was supported under Australian Research Council's Discovery Projects funding scheme (project number FL170100160).

Abstract

Fitness is a central but notoriously vexing concept in evolutionary biology. The propensity interpretation of fitness is often regarded as the least problematic account for fitness. It ties an individual's fitness to a probabilistic capacity to produce offspring. Fitness has a clear causal role in evolutionary dynamics under this account. Nevertheless, the propensity interpretation faces its share of problems. We discuss three of these. We first show that a single scalar value is an incomplete summary of a propensity. Second, we argue that the widespread method of "abstracting away" environmental idiosyncrasies by averaging over reproductive output in different environments is not a valid approach when environmental changes are irreversible. Third, we point out that expanding the range of applicability for fitness measures by averaging over more environments or longer time scales (so as to ensure environmental reversibility) reduces one's ability to distinguish selectively relevant differences among individuals because of mutation and eco-evolutionary feedbacks. This series of problems leads us to conclude that a general value of fitness that is both explanatory and predictive cannot be attained. We advocate for the use of propensity-compatible methods, such as adaptive dynamics, which can accommodate these difficulties.

KEYWORDS

eco-evolutionary feedbacks, environment, expected reproductive output, fitness, propensity

INTRODUCTION

The concept of fitness is central to evolutionary biology. Without fitness differences between individuals in a population, there would be neither natural selection nor adaptation (i.e., cumulative adaptive evolution). One might accordingly think that evolutionary biologists have settled on the definition for such a fundamental concept. If only that were so. To this day, fitness is defined in different and even inconsistent ways.[1–15] In light of this, it is not at all uncommon for prominent evolutionary biologists to concede that "Unfortunately, fitness is difficult to define more specifically so that it can be measured and understood more clearly."[16] Routine concessions like this suggest that little has changed since Stephen Stearns proposed the following satirical definition: "Fitness: something that everyone understands but no one can define precisely."[17] Such quips and concessions belie a fundamental challenge: If evolutionary theorists are entitled to deploy inconsistent definitions and corresponding measures of fitness on pragmatic grounds, then it becomes possible for two (or more) evolutionary biologists who observe an evolving population and have all the relevant information about the system in hand to reach incompatible conclusions about whether or to what extent natural selection occurs.[18] We find such a possibility deeply disconcerting and suspect that others share this unease. In this paper, we provide a rationale for constraining the meaning(s) of this concept by demonstrating intrinsic limitations to its measurement. Fitness can prove a useful if not indispensable concept when a target population and its environment meet several nontrivial conditions; namely, that comparisons involving two or more competing trait types (or individuals) cannot be made if there are irreversible changes in the environment. However, as we shall show, these conditions can be satisfied only in much restricted spatio-temporal partitions of the total (biotic and abiotic) environment. A consequence is that fitness cannot be rightly regarded as a general measure of or "forecast for" an organism's long-term evolutionary success.

FITNESS IS NOT ACTUAL REPRODUCTIVE OUTPUT

Nearly everyone accepts that fitness is associated with a notion of viability that eventuates in reproductive output. A correspondingly intuitive proposal would then be that fitness is nothing more than an individual's actual reproductive output. If, however, one uses actual reproductive output as a definition of fitness, natural selection becomes an empirically unfalsifiable (tautological) concept.[1,19,20] The fittest individuals in a population just happen to be those that survive and produce the most offspring irrespective of the reasons for their having done so. Consequently, greater fecundity would no longer be reliable evidence upon which to infer the character states that are better able to meet ecological challenges to survival. As biologists have long known,[21–23] an individual's actual reproductive output should provide evidence of, while not being an exhaustive definition for, its fitness.

While the foregoing definition of fitness is easily dismissed, another definition must stand in its place. Philosophers of biology have proposed a definition that characterizes fitness as a "propensity" to survive and produce offspring.[24–29] Box 1 presents the philosophical context in which the propensity interpretation arose as well as some of the classical difficulties associated with it. On the propensity view, two individuals with distinct character states or phenotypic variants can have different probabilistic dispositions to produce certain numbers of offspring. An individual's actual reproductive contribution is evidence for the expected fecundity of its character state type, which can be compared against extant competing character state types to yield measures of relative or differential fitness. Assuming a uniform environmental background, we can accordingly predict that any individual bearing the character state with higher expected fecundity will be favored by natural selection. We can also account for the fact that individuals with the optimal character state sometimes fail to leave the most offspring since there can always be highly localized environmental fluctuations that prevent an individual from realizing its capacity to contribute the expected number of offspring for its trait type. By way of a crude analogy, assume that vases have different propensities to break when dropped. Despite some types of vase being less fragile on average than others, it would not be inconceivable or even all that surprising if, from time to time, a single vase of a more fragile type did not break when dropped on the floor.

An oft-touted virtue of the propensity interpretation of fitness is that it happens to coincide with a probabilistic account of reproductive outputs. Fitness can accordingly be translated into and from a probability distribution (Figure 1). This welcome feature is often taken as suggesting that there is nothing more to the definition of fitness than the mathematical expectation, expressed as a scalar numerical value, used to denote it.[26] But fitness as a propensity makes it a physical property of an individual. It is supposed to be an explanatory (causal) property of an individual organism rather than an individual type, albeit one that does not admit of direct empirical access and must instead be inferred via statistical means.[1] At first pass, this certainly sounds perplexing since fitness is a quintessentially relational property. The degree of an organism's "adaptedness to" or "fit for" a particular set of environmental challenges is what enables it to survive and reproduce in that type of environment. There is no such thing as fitness *simpliciter*. Propensity theorists do not deny this. Their contention is that fitness ascriptions need not make explicit reference to the specificities of the environment.[27] The reference environment, on the propensity interpretation, potentially consists of all possible background conditions that could in principle be sampled by competing trait variants in the long run. In the extreme, it could be construed as unchangeably broad in scope, a "total" environment that presumably functions as a prerequisite for genuine fitness differences that both explain a causal history of competitive success and predict future representation. Any further environmental specification would consequently generate a special case of this all-encompassing reference environment.

Following the propensity interpretation of fitness, natural selection is then the process that leads individuals with the "highest propensity to leave offspring" (i.e., those that exemplify the trait type combinations with the greatest expected fecundity) to increase in relative frequency. When individuals with the highest propensity leave fewer offspring than expected, individuals with the lowest propensity leave more offspring than expected, or individuals with the same propensity leave different numbers of offspring, the evolutionary process at work is not natural selection but drift. Drift is typically derived on the basis of departure from expected reproductive outputs (Box 2b). It measures the extent to which evolutionary change (or the lack thereof) results from so-called "accidents," or the causal factors that filter extant heritable variation in a population indiscriminately.[63,65–69]

As definitions of fitness go, the propensity interpretation has been widely acclaimed by philosophers of biology. Much of the support it has drawn is due to its apparently seamless accommodation of the statistics and probability theory that biologists deploy to measure fitness. We now introduce three difficulties that arise directly from the mathematical methods on which the propensity interpretation depends.

PROBLEM 1: THE FAILURE OF REPRODUCTIVE EXPECTANCY

Even though depicted as having an objective physical reality on the classical propensity interpretation of probability,[70] propensities differ from other quantities like mass or size in that there is no scale or unit of measure for them. Propensities, in general, do not admit of direct measurement. One can only access the outcomes of trials (e.g., the side on which a coin falls, whether the vase is broken, or the actual reproductive output of an individual) as well as some of the physical properties of the system generating them (e.g., the shape of the coin, the composition of the vase, or the traits of the individual). Assigning numerical values to a propensity, then, is necessarily an indirect inference. This inference depends on statistically combining sequences

BOX 1: Philosophical Context

Any analysis of fitness should account for two features of this concept. On one hand, fitness is invoked as a causal parameter in explanations for why some character states or trait variants (and the organisms that exemplify them) are more prevalent than competing character states or trait variants in a population. On the other hand, adequate measures of fitness should provide the means by which to accurately predict the direction and magnitude of relative frequency changes due to natural selection. While these can be seen as distinct projects (i.e, definitional analysis versus mathematical measurement),[30] philosophical discourse has nevertheless concentrated on whether an explication of the concept can satisfactorily capture both features.

Philosophical debate about fitness comes to the fore with criticisms lodged by Smart,[31] Manser,[32] and perhaps most of all, Popper.[19] Popper, for instance, noted that if the fittest are just those that survive and reproduce in greater number, then evolutionary theory is unfalsifiable because the central notion of fitness is tautological. If the "fittest" are defined as, for example, "those that produce the most offspring," then those that happen to produce the most offspring must always be deemed the fittest without regard for whether that reproductive success is in fact due to their actually being better adapted to the challenges posed by the selective environment. Much the same problem had already been laid bare by Scriven,[20] who, using an example involving identical twins who exhibit different viability, shows why *actual* or *realized* fitness cannot suffice as a definition of fitness even if it is occasionally an adequate estimate of fitness. Such difficulties prompted some, Williams[33] and Rosenberg[34–36] in particular, to argue that fitness is best construed as an undefined theoretical primitive. At the time, however, most philosophers of biology found this suggestion wanting.

Brandon[24,37] and Beatty[26] recognized that any adequate explication of fitness had to be empirically sensitive to actual reproductive output, which is, of course, the evidentiary basis for fitness estimates, without being exhausted by actual lifetime fecundity or viability. Beatty's and Brandon's proposals shared two key components. The metaphysical or ontological component depicts fitness as a probabilistic dispositional property or "propensity." Fitness is thereby an intrinsic and objective feature of a token organism, albeit one that typically finds expression only in the organism's relation to a specified selective environment. The epistemological component acknowledges the mathematical means by which to measure this propensity. It is estimated via the statistical expectation for a character state, or a probability weighted average over reproductive outcomes for a trait variant after a specified period.

This was seen by many as a major insight. It permitted individual organisms exhibiting a character state to deviate from the statistical expectation for that type, while maintaining the capacity to explain why on average some character states did better than others in a homogeneous selective environment. And it apparently provided firm causal and explanatory footing for fitness as a property of individual organisms, the basic constituents of the populations whose dynamics are targeted by evolutionary explanation. Moreover, it was faithful to the practice and methods of biology.

Problems nevertheless arose for the propensity interpretation. The most pressing difficulties can be traced back to work by the theoreticians Thoday, Gillespie, Levins, and Lewontin.[38–43] With an appreciation of these contributions, Beatty, Brandon, and Sober[25,44,45] each showed how assigning a single, unchanging numerical measure of fitness could generate an erroneous estimate for the fitness of a character state and, therefore, relative fitness differences. Equating the expectation with arithmetic mean offspring contribution, for example, clearly failed to take account of crucial information about intra- and intergenerational reproductive variance among competing character states. Higher mathematical moments of statistical distributions (e.g., skew, kurtosis, etc.) pertaining to offspring output might likewise generate inconsistencies. Nor could these inadequacies be casually dismissed as mere mathematical artefacts. The mathematical models unambiguously represented biologically real and important phenomena; namely, (1) demographic stochasticity (i.e., within and between generation differences in number or timing of offspring) and (2) environmental stochasticity (i.e., fluctuations in the biotic and abiotic components of the selective environment).

Without sufficient means for measuring fitness, worries about the already contentious metaphysical status of fitness as a propensity motivated some to dispense with the propensity interpretation altogether. By way of example, advocates of the so-called "(merely) statistical interpretation" of fitness argue that fitness is predictive but neither causal nor explanatory.[46–51] On this view, fitness estimates do not correspond to the intrinsic, inheritable causal basis of the success or failure of individuals. Fitness measures are instead post hoc redescriptions over a virtually unlimited number of token causal events that influence survival and reproductive outcomes in finite populations. Others who are no less sceptical of propensities, such as Abrams,[52–54] have proposed more moderate alternatives that do not jettison the explanatory ambitions of the concept.

Despite several spirited rejoinders,[55,56] those who sought to defend a causal interpretation of fitness were somewhat at a loss until a new mathematical foundation for the propensity interpretation was developed by Pence and Ramsey.[27] Working with Ramsey's notion of "Block fitness,"[28] they drew on a mathematical framework known as "adaptive dynamics."[8,57] Adaptive dynamics is an extension of evolutionary game theory to dynamic ecological scenarios that rely on feedbacks. Doing so enabled them to resolve several of the outstanding difficulties associated with demographic and environmental stochasticity. The fitness function derived by Pence and Ramsey

BOX 1: Philosophical Context

generates a static scalar value of fitness (*F*) for any individual in a population by exploiting a massive multidimensional state space that they call "Ω." It consists of all the possible lineages to which a token organism, identified via its genotype, might give rise. This function can preserve a static value of individual fitness in the face of many challenging and ubiquitous selective scenarios when certain nontrivial provisions (see Sections 4 and 5) are met. For a more comprehensive overview and discussion of fitness in the philosophy of biology, see Rosenberg and Bouchard.[15]

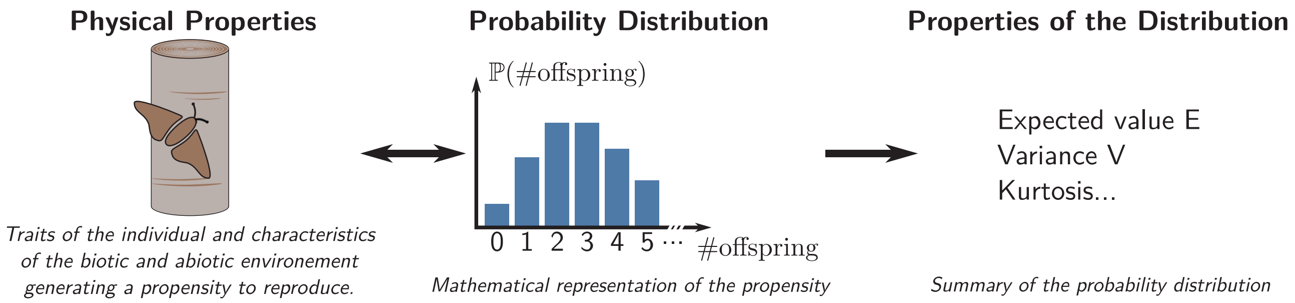
Physical Properties
Probability Distribution
Properties of the Distribution
$\mathbb{P}$(#offspring)
0 1 2 3 4 5 ···
#offspring
Expected value E
Variance V
Kurtosis...
Traits of the individual and characteristics of the biotic and abiotic environement generating a propensity to reproduce.
Mathematical representation of the propensity
Summary of the probability distribution

FIGURE 1 Propensity to produce offspring is a physical property of the relation between an individual and its environment. The physical characteristics of an individual and its environment generate a propensity to produce offspring. Just as a force like weight (a physical relational property of an object and the earth) is represented mathematically by a vector, a propensity can be represented mathematically by a probability distribution, such as the probability of leaving *x* offspring during the lifetime of the individual. A probability distribution can be summarised by a few key properties (such as its moments, i.e., means, variance, kurtosis), but this characterization is generally incomplete

of outcomes or insight into the physical structure of the system.[71] Unfortunately, the exact relationship between the physical properties of individuals and their respective reproductive outputs is currently beyond the ken of biology, perhaps with the exception of what transpires in the simplest of ecological and laboratory scenarios.[72] In practice, fitness is thus often condemned to be inferred merely on the basis of actual reproductive output.

A propensity is often represented by a single privileged number, namely its expected value. The expected value of a probability distribution is defined as the average of outcome values weighted by their probability. It is often preferred over other scalar descriptors (e.g., median, maximum, variance, kurtosis, or a combination of these) as a first approximation of an outcome distribution because of properties that follow directly from the axioms of probability theory. The "law of large numbers" is particularly important in this regard. It states that the average of independent samples from a distribution will converge toward its expected value. This consequence is independent of the philosophical justifications (epistemological vs. metaphysical) for applying the axioms of probability. The expected value correspondingly plays a central role in both statistical inference (e.g., parameter estimation) and projection.[73,74]

But identifying a propensity with its expected value can be just as misleading as identifying fitness with realized reproductive output (as discussed in the previous section). When considering simple dichotomous events, identifying a propensity with its expected value is not only intuitive but sensible. This is so because there is little loss of information: the whole distribution can be perfectly described by a single number. By way of illustration, the propensity interpretation of the fairness of a coin correspondingly reduces this probabilistic quality to a unique expected value: 0.5 (provided that we assign to each side the numerical value 0 and 1). The expected outcome must be 0.5 in order for a coin to be deemed fair. Contrast this with the propensity interpretation of fairness for a six-sided die (Fig. 2). The fairness of this die cannot, in similar fashion, be reduced to an expected value of 3.5. For even a die that can fall on nothing other than one and six yields this expected value. Such a die, however, is conventionally unfair or "loaded." Propensity and expected value can thus come apart. The expected value only defines a probability distribution uniquely when considering dichotomous outcomes or when the shape of the distribution is constrained by a one-parameter model. Unfortunately for the problem of fitness, an individual's reproductive output—number of offspring—is neither dichotomous nor constrained by a unique model. Therefore, it cannot be reduced to a single number.

The foregoing discussion demonstrates that expected reproductive output should not be hastily equated with fitness as a propensity to reproduce. It can misleadingly conceal the complexity of a probability distribution.[25] An expectation may nevertheless be used as a legitimate, meaningful, and simple summary of this propensity. In the following section, we direct attention to cases in which such a single-valued, scalar measure proves to be a reliable quantification of individual reproductive propensity for causal explanations of evolutionary dynamics. This will show that there is much worth retaining in the propensity interpretation of fitness even if some of its ontological commitments are deemed extravagant.

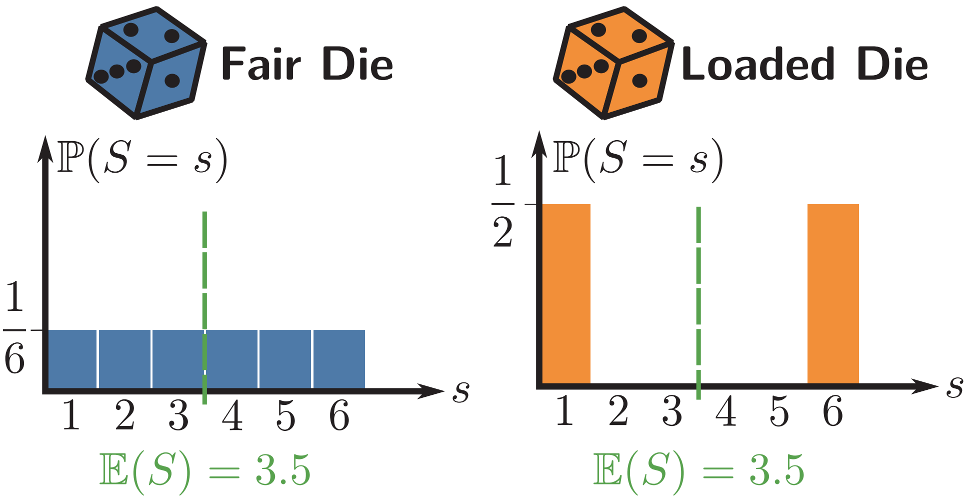

FIGURE 2 The expected value does not capture the full propensity. In this example, both dice have the same expected outcome (Score *S* of 3.5), but only the blue one is conventionally considered fair, while the second is "loaded". The die's propensity to be fair cannot be summarized by a single scalar value. See main text for details

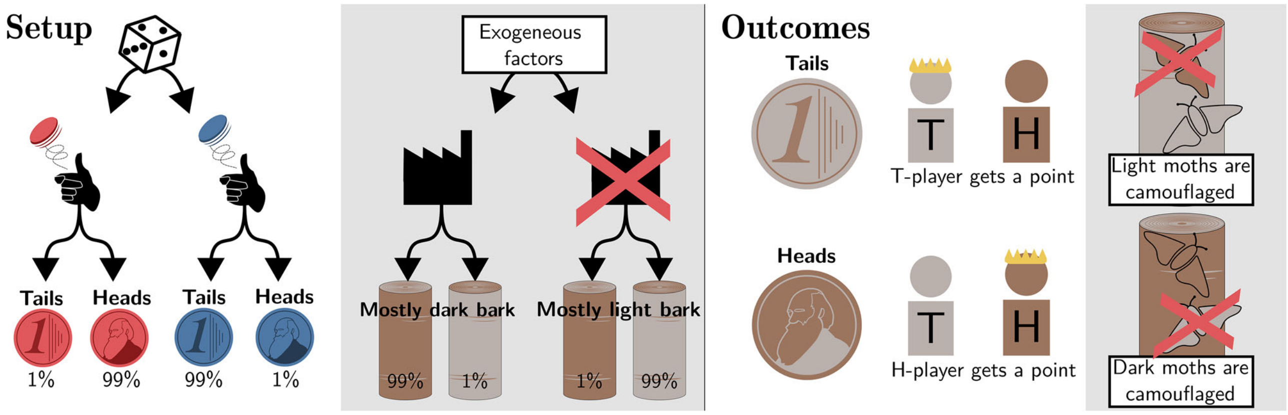

FIGURE 3 Die-and-Coins Game. In this game, a player earns a point if the outcome of the coin toss matches their bet (heads or tails). The game is played with two biased coins (red and blue). The coin actually used for play is determined by casting a die. This game is explicitly designed as an analogy for a simple ecoevolutionary scenario wherein each player is a peppered moth with a given pigmentation (melanic or light). A moth gains an advantage (survival due to camouflage from predators) only if they land on similarly coloured bark. The forest is either overwhelmingly composed of light or dark birch trees depending on an exogenous factor such as the presence of a factory. See main text for details

PROBLEM 2: ENVIRONMENTAL AVERAGING CAN LIMIT PREDICTIVE EFFICACY

Fitness measures should facilitate predictions concerning the fate of evolutionary systems, such as the frequency changes for individuals of a given type. This is typically achieved by computing a projection of the future population size from the long-term reproductive output of individuals.[75] Yet, as demonstrated in the previous section, reducing a probability distribution to a single scalar value only suffices in the most simple dichotomous cases. Despite this difficulty, fitness is still often deployed in the form of a single number, typically the *arithmetic mean* of the exponential growth rate of a type in a population across possible environments. The fittest type or individual is accordingly the one with the highest mean exponential growth rate. In Box 2a, we explain why this value is mathematically equivalent to using the *geometric mean* of expected reproductive output in discrete-time models like those that feature in the philosophical literature.[44,45] But the (exponential) growth rate used for projection must hold true for the period of time under consideration if it is to render an accurate prediction. Problems can arise if the environment changes over this time period in a way that affects an individual's long-term reproductive output.[76,77] One way to sidestep this complication is by computing the exponential growth rate in an idealized "average" environment (i.e., averaging over all the environments encountered by the individuals, weighted by their relative frequency). In this section, we show how irreversible changes in the environment can preclude this approach.

To see why this is so, consider a simple game played with a die and two coins (Figure 3). The two coins are biased in such a way that the

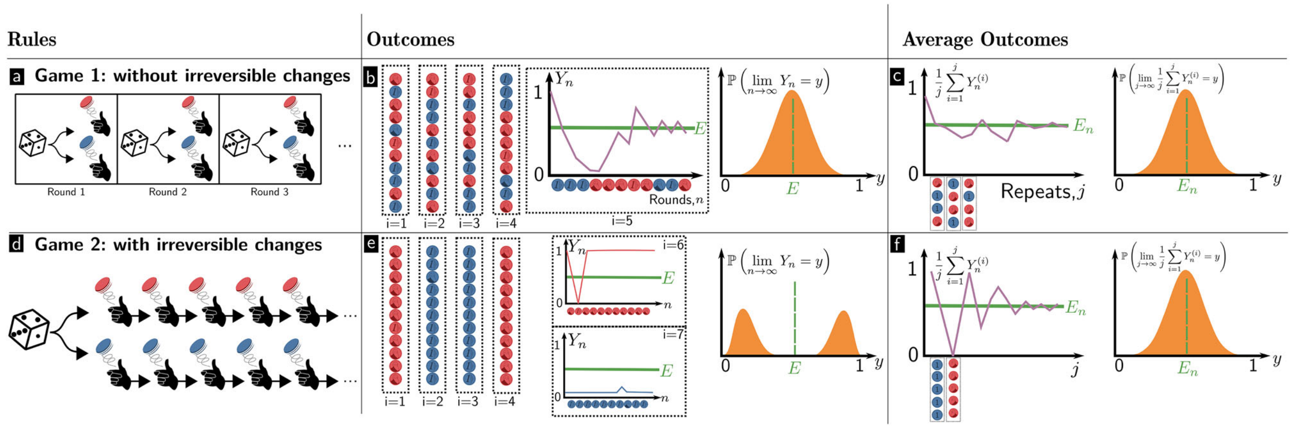
Rules
Outcomes
Average Outcomes
a Game 1: without irreversible changes
Round 1
Round 2
Round 3
...
b
1
Y_n
E
0
Rounds, n
i=1
i=2
i=3
i=4
i=5
$\mathbb{P}\left(\lim_{n\to\infty} Y_n = y\right)$
0
E
1
y
c
1
$\frac{1}{j}\sum_{i=1}^{j} Y_n^{(i)}$
E_n
0
Repeats, j
$\mathbb{P}\left(\lim_{j\to\infty} \frac{1}{j}\sum_{i=1}^{j} Y_n^{(i)} = y\right)$
0
E_n
1
y
d Game 2: with irreversible changes
...
e
i=1
i=2
i=3
i=4
i=6
i=7
1
Y_n
E
0
n
$\mathbb{P}\left(\lim_{n\to\infty} Y_n = y\right)$
0
E
1
y
f
1
$\frac{1}{j}\sum_{i=1}^{j} Y_n^{(i)}$
E_n
0
j
$\mathbb{P}\left(\lim_{j\to\infty} \frac{1}{j}\sum_{i=1}^{j} Y_n^{(i)} = y\right)$
0
E_n
1
y

FIGURE 4 Two variants of the die-and-coins game. Game 1: (a) The die is cast at the beginning of each round, so there are no irreversible changes in the color of the tossed coin. (b) The future long term average score of a player (i.e., the limit for increasing n values of Y^n, the average score of a player at round n) does not depend on the specifics of the present. And the actual average score always converges toward its expected value E. (c) Independent repeats of Game 1 can be predicted using the law of large numbers. For any given round n, the average outcome of many independent repeats $Y_n{}^{(1)}, Y_n{}^{(2)}, \ldots$ converge toward the expected value E_n of the n-th round. Game 2: (d) The die is cast once at the beginning of the game. There is an irreversible change in the colour of the tossed coin. (e) The future long term average score of a player depends on the specifics of the present and does not converge toward its expected value E. (f) Independent repeats of Game 2, like with Game 1, can be predicted using the law of large numbers. An interactive version of this figure is available assupplementary material

red coin almost always falls heads while the blue coin falls predominantly tails. The die is used to decide which of the two coins will be tossed. Each player is assigned a side of the coin at the beginning of the game and earns a point each time the outcome of a toss matches their respectively assigned side. We will accordingly refer to the "heads player" ("tails player") as the player who earns points whenever heads (tails) is revealed.

This game is designed to parallel the classical evolutionary scenario involving industrial melanism in peppered moths (see caption of Figure 3).[78,79] Dark and light phenotypes are equivalent to heads and tails players in this game, while the colour of the coin that is tossed refers to the nature of the environment in which the moths live. A lighter birch tree forest overwhelmingly favors the light moths that are camouflaged on the light bark, just as the red coin gives an advantage to the heads player. Each coin toss represents an event in the life of a peppered moth that affects its reproductive output (e.g., encountering a predator). Earning a point in the game should be interpreted as reaping the benefit of a favourable event (e.g., evading a predator) . The die represents exogenous factors that constrain the environment, such as the presence or absence of a nearby factory.

Just as the lifetime reproductive output of an individual is the result of many events, this game is one that is decided over many rounds. In the following, let X_n be the score of the heads player at round n and let $Y_n = \frac{X_n}{n}$ be the average score per round for this player. The experimental arrangement of the game (i.e. the physical characteristics of the die and coins, along with the rules of play) generates a propensity for this score (Y_n), one that is captured by a probability distribution that maps each possible score value of Y_n to a specific probability value [the set of $P(Y_n = y)$ for all values of y].

First, consider a set of rules (Game 1, a) in which the die is cast at each round, leading to a potential change in the colour of the coin at each round. For instance, assume that the red coin will be used whenever the die shows an even number, while the blue coin will be used whenever the die shows an odd number. Provided that the die is fair and that the coins are biased in equal but opposite ways (i.e., symmetrically), the expected value of the average score per round for a player is 0.5. During a game, the outcome of a particular round n, X_n , can be considered an independent sample of a single probability distribution D. By the law of large numbers, the long-term average score of the player, Y_n, converges toward E, the expected value of D, which is 0.5 (Figure 4b). Note that this prediction is true no matter what the outcome of the first toss. From the biological standpoint, this fitness value is adequate no matter what the initial state of the forest.

This may seem like a highly unrealistic setting, as the environment at one time is independent from the environment at a previous time: the color of bark in the forest may completely change between rounds. Metaphorically, the environment has thus far been characterized as if it had no memory of its prior state(s). A less artificial picture would introduce historicity in the environment. Let us entertain this possibility by considering another version of Game 1 in which the rounds are not independent (not represented in Figure 4). This would be the case if, instead of allowing the die to determine which coin is used at each round, the coin used between two rounds remains the same unless the result of a die roll is six, in which case the coin is changed (from red to blue or vice versa). In such a situation, a unique scalar value that accurately predicts the long-term average score irrespective of the present state can be obtained, but only if coin-color change (i.e., reversibility) remains an open possibility (i.e., there is a non-zero probability of change). However, the law of large numbers alone no longer suffices to justify this claim. It must be supplemented with results from Markov chain theory.[2] The mathematical subtleties of this supplementary theory need not distract us from the crucial point here: if there are no irreversible changes in the environment, the average score per round converges in time toward a unique value, no matter what the specifics of the present (i.e., the first few tosses). This single value is the average score of the player, weighted by the probability that a given coin is used in the long term, a value that is in turn determined by the actual bias of the die (i.e., given by the stationary distribution of the Markov chain). It is independent of the initial conditions (i.e., whether the game started with a blue or red coin). Provided one knows the relative probability of the forest being dark or light in the long run and that the forest never changes irreversibly, one can accurately predict the long term success of a moth regardless of the current state of the forest.

Now, consider a new set of rules (Game 2, Figure 4d) under which the die is cast just once before the first round. The color of the coin never changes in this version of the game. This mirrors a scenario in which there is irreversible change to the environment, as might be the case if a light birch forest was forever darkened once a nearby factory opened. Since the coin-color remains constant for the duration of an entire game, the long-term average score for a given player cannot be accurately predicted by averaging over the two exclusive sets of trials involving *only* red coins or *only* blue coins.[3] If the forest were to become irremediably dark, averaging the success of moths over all possible environments (i.e., both dark and light forests) would likewise only worsen the prediction by including information about potential rates of success in unreachable light environments (Figure 4e). In this scenario, no unique scalar value can discount the specifics of the present and accurately predict the future. In Box 2b we explore the links between Game 2 and the notion of drift.

Note here that if one were to repeat Game 2 several times (indexing the independent trials by $i = 1, 2, \ldots$, like so$Y_n^{(1)}, Y_n^{(2)}, ...$) and, then, take an average of the average score per round computed at round *n*, the value (the average of the$Y_n^{(i)}$ for all values of *i*) would always converge toward a value E_n as *i* increases (Figure 4C, 4f). Note also that E_n converges toward E as$n$ increases. This higher-order form of convergence is an immediate consequence of the independence of the games and hinges on the law of large numbers. But interpreting a biological scenario as conforming to this type of convergence comes with pitfalls. Prominent among these is that doing so requires, as just mentioned, assuming complete independence between repeats of a game and thus the absence of irreversible changes in the environment. The previously discussed limitations pertaining to the independent variant of Game 1 accordingly resurface.

The various die-and-coin games discussed above, as well as their biological equivalents, lay bare an important caveat for any attempt to predict the outcome of a dynamic process via a single scalar value. Namely, these show why a fitness value must assume the absence of irreversible changes in the environment (see Figure 4a–c) or else ensure its own inadequacy (see Figure 4d–f). Yet, as we shall show in the next section, most environmental scenarios appear to be approximately irreversible. This presents a daunting challenge that threatens to undermine any general measure of fitness.

PROBLEM 3: MUTATION-INDUCED TRADEOFFS BETWEEN SCOPE AND RESOLUTION

The criticism lodged in the previous section should not be taken as sounding the death knell for fitness measures that average over all possible environments. It merely shows that under some conditions, such as for irreversible environmental changes (Game 2, Figure 4d), accurate prediction is unachievable. But if we assume that, at each time step, the factory has a nonzero but very small chance of either closing if it was present (i.e., lightening the dark trees) or appearing and opening if it was absent (i.e., darkening the light trees), then the average growth rate value over all possible environments considered constitutes a fitness measure that holds true *in the long run* (Figure 5a).

There is, however, a major drawback with this approach, one that stems from the fact that organisms can mutate. Recall that a desideratum for any scalar fitness measure is that it should both predict the evolutionary dynamics of a population and permit comparing the success of different individuals or phenotypes. Consider, again, the example involving peppered moths in this context. Achieving as much would require comparing the fitnesses of types in all possible environments (i.e., abstracting away the color of the bark altogether), which in turn presumes that we already know the respective reproductive outputs of those types in each environment of interest (on light vs. dark bark). Doing this in a way that maintains predictive accuracy compels us to examine a timescale at which the colour of the bark is reversible and, then, compute a weighted average of fitnesses over that timescale. Such a situation is depicted in Figure 5a. In this figure, the timescale under consideration ensures environmental reversibility. The factory is present, then absent, and finally present again, while corresponding changes to the colour of bark ensue. There is no mutation in this scenario; surviving moths always breed true to form. Consequently, in the long run, the average growth rate of each type of moth converges toward a value that averages their growth rate on light and dark barks weighted by the proportion of time spent on each color of bark (just as Y_n converges toward E in Game 1).

By explicit contrast, mutation between types is a possibility for the scenario shown in Figure 5b. This scenario is otherwise identical to that depicted in Figure 5a. Over a timescale at which both environments are accessible, it is reasonable to assume that some descendants of any focal individual would mutate so as to express the alternate phenotype. Even if we assume a high level of fidelity between generations and stipulate that there is no extinction, the initial phenotype of an individual becomes increasingly irrelevant as the number of generations considered increases. Assuming the possibility of mutation, the long-term number of descendants becomes a function of all the phenotypes that are potentially accessible via mutation within all of the potentially accessible environments. Since the effects of both initial phenotype(s) and initial environment(s) become diluted in the long run, any fitness value obtained by this approach maintains its predictive integrity only by glossing over the fitness differences between the competing types. Consequently, in the long run, the average growth rates for both types of moth converge toward the same value.

BOX 2: Arithmetic Mean, Geometric Mean, and Drift

This box discusses two topics related to the concept of fitness that have featured in the philosophical literature, namely the use of geometric mean as opposed to arithmetic mean to compute fitness and the concept of drift.

a) Fitness: Geometric or Arithmetic mean?

As argued in Section 3, it is impossible to exhaust the information of a probability distribution more complex than a Bernoulli trial (i.e., with a binary outcome) with the mean of a single scalar value. While it is impossible to capture the entirety of the distribution, it is possible under some assumptions to accurately capture a relevant behaviour of the lineage spawned by the individual with a single scalar (provided that descendants behave similarly to the ancestor). Often the long run growth rate of the logarithm of the population is used.[27,57]

Why use the growth rate of the *log-population* and not the growth rate of the population ? The reason is that populations do not grow linearly but geometrically (or exponentially if modelled in continuous time). There is accordingly no way to define a growth rate of the population as the slope of a line. Although rarely explicit, what most practitioners actually refer to when using expressions like “growth rate of the population” is the growth rate of the *log-population*. This ambiguity also accounts for the occasional reference to the geometric rather than arithmetic mean in the philosophical literature.[45]

A numerical example much like the one provided by Beatty and Finsen makes this point evident.[44] If an individual of a particular type produces two offspring and each of these offspring produce four grand-offspring, this is equivalent to a situation in which individuals produce $(2 \times 4\ )^{1/2} = \sqrt{8} = 2.83$ offspring each generation, not $\frac{2+4}{2} = 3$offspring. This becomes obvious if one notes that the progenitor has eight grand-offspring, as if the population was multiplied by the geometric mean $\sqrt{8}$ each generation (e.g. $\sqrt{8} \times \sqrt{8} = 8$) rather than the arithmetic mean 3 (e.g. $3 \times 3 = 9$). Consequently, the proper way to *average* a geometric growth rate is to take the *geometric mean* of the multiplicative terms. In log-space, however, the *arithmetic mean* gives the correct result: the log of the population increases by $\frac{ln(2)\,+\,ln(4)}{2}$ each generation (for a total increase of $ln(8)$).

This result should not be surprising since fitness computed using the arithmetic mean of the log-population increase is strictly equivalent to computing it with the geometric mean of the population multiplicative terms in the sense that $e^{\frac{ln(2)+ln(4)}{2}} = (2 \times 4)^{1/2} = \sqrt{8}$ by the basic properties of the log function. This is always true for multiplicative processes regardless of their deterministic or stochastic nature: the arithmetic mean of the exponential growth rate is equivalent to the geometric mean of the multiplicative growth rate. However, in the absence of intergenerational fluctuation (i.e., when the population is multiplied by m every generation), the arithmetic and geometric means are equal $(\prod_{i=1}^{n} m)^{1/n} = \frac{1}{n} \sum_{i=1}^{n} m = m$, so this distinction is unnecessary.

b) Drift and Expected Values

Drift is an elusive concept in evolutionary theory as it refers to different phenomena that have to do with chance.[81] It has been proposed as a causal process as well as an outcome.[82] A classical way to represent drift is to think about identical twins exhibiting the same trait type, one of whom is struck by lightning while the other survives to produce offspring.[20,83] “Accidents” like this show that an individual’s actual reproductive output can diverge from the expected output for its type. Such an instance of drift is an example of *sampling* drift.[84]

The random sampling of individuals induces a departure from the expected outcome. By way of analogy with Game 1, a case of sampling drift would occur when a player experiences an unexpected sequence of heads or tails. Sampling drift has a particularly strong effect in small populations, just as the effect of an individual struck by lightning in a population consisting of four individuals is more noticeable than it would be were it in a population of one-thousand. It is largely for this reason that random drift is of prime importance in situations known as population “bottlenecks” (e.g., founder’s effect) but typically negligible in large populations.

Game 2 also shows a deviation from expectation: the actual score of any player never converges on its expected score. Such an event is random insofar as it depends on the roll of a fair die and can arguably be classified as a form of drift. However, the mechanism producing drift here is qualitatively different from the sampling form of drift discussed above. The effect of the factory’s closure (initial roll of the die) on the evolutionary trajectory of the moth variants is *not* negligible even in a large population. The fact that the die is rolled just once explains why trajectories are unrepresentative of the die’s expectation. Moth population size is not explanatorily relevant here. This source of random drift can nonetheless be compared to other sources of stochasticity by referring to the *effective* population size. Given a deviation from expectation stemming from an unknown “drifty” process (e.g., variance in allele frequency), the effective population size represents the census size that a population adhering to the assumptions of a model in which there is only sampling drift (e.g., the Wright-Fisher model,[23] but other models could be used) would have if it were to display the same deviation from expectation. Increased environmental noise usually results in reduced effective population size. But it is important to note that relying on effective population size to characterize the level of drift in a population does not provide any insight into the mechanism(s) producing the deviations from expectation.

BOX 2: Arithmetic Mean, Geometric Mean, and Drift

The foregoing implies that census population size and drift can be decoupled when attempting to give a mechanistic description of drift. This point, with a few exceptions,[63,69,85] has too often been missed in the philosophical literature. It can also be made from an individual organismic perspective. Any source of environmental heterogeneity can increase the variance in a given type's propensity to reproduce by decreasing the chance that the individuals constituting the type realize the "lives" (reproductive outcomes) that they do. Grant Ramsey calls this property of individuals "driftability."[69]

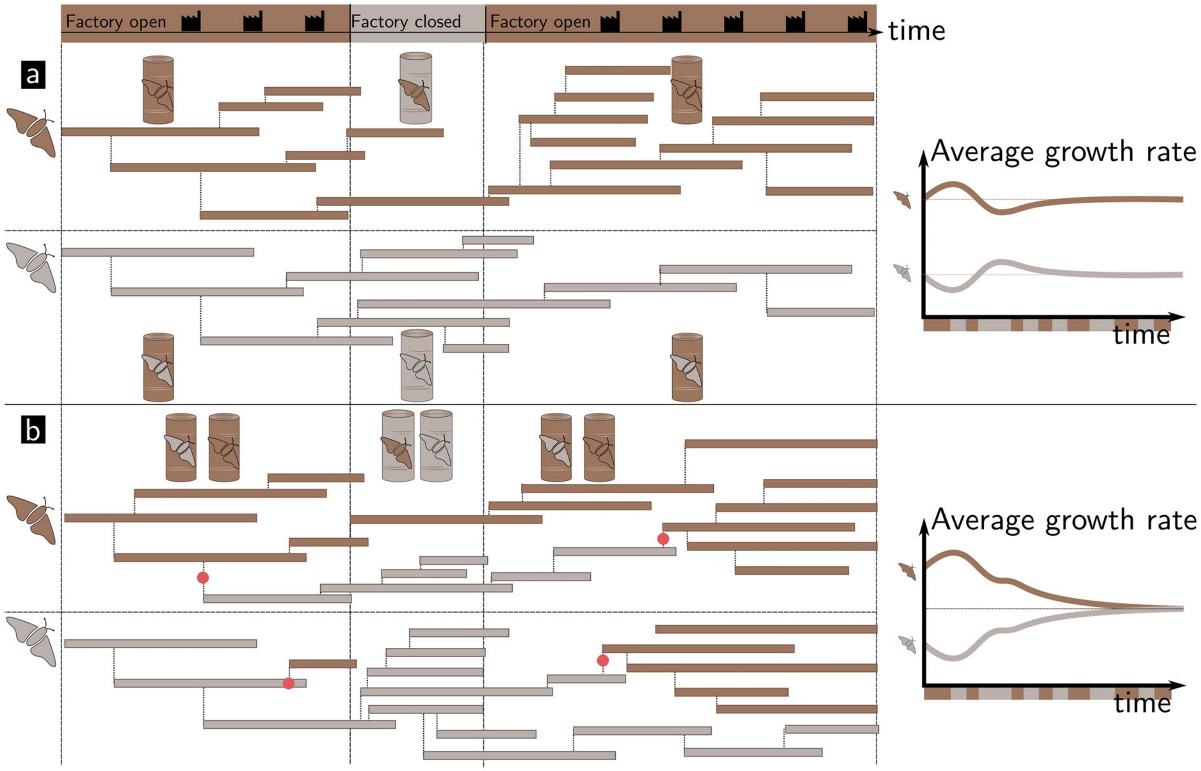

FIGURE 5 Considering longer-term outcomes warrants averaging over more environments, but reduces resolution because of mutations. Horizontal lines represent the lifespan of an individual and are colored according to their phenotype. Vertical lines represent reproduction. The x-axis is time, the y-axis has no specific meaning other than preventing overlapping. Graphs on the right show the average growth rate computed over a variable period of time, similarly to Yn in Fig. 4b and 4e. (a) No irreversible change in the environment, irreversible phenotype. The average growth rates of each type converge toward unique values (horizontal lines) that reflect the average environment and are independent of the initial environment. (b) No irreversible changes in the environment, no irreversible changes in phenotype (red circles represent back mutations that change the phenotype of the moth). The average growth rates of both types converge toward a unique value (horizontal line) that reflects the average environment and is independent of both the initial environment and the initial type

The foregoing example reveals a fundamental conceptual trade-off. A broad-scope fitness measure—namely one that averages over multiple environments—is predictive only at a timescale over which all of these environments can be reached. This is because, in the long run, the initial environment weighs less on the outcome compared to the average reachable environment. But mutations change the type of individuals in the long run, so the initial phenotype weighs less on the outcome than the average phenotype. The phenotypic distribution of descendants subsequently becomes indistinguishable in the long run. If a measure of fitness is to be both predictive and capable of distinguishing the evolutionary trajectories of types, then it must consequently be defined at a timescale over which two conditions hold: (1) All considered environments must be potentially accessible and (2) the phenotypic states in question must *not* be inter-accessible via mutation. If the first condition goes unmet, predictive efficacy cannot be achieved. If the second condition is violated, resolution at the timescale for which predictive efficacy is preserved becomes so low that relative fitness differences between individuals or types disappear. This trade-off between predictive accuracy and differential fitness is especially pressing when attempting to define general fitness values that average over a large diversity of environments and ignore heredity, as some authors have proposed.[86]

The choice of a reference environment, including the relevant timescale, is thus central to any definition of fitness that promises to inform its measurement.[87] It has been recognized in the literature that both a “long” and “short” term concept of fitness could coexist concurrently.[30,44,45,53] The examples introduced in previous sections allow us to show precisely why this is so: a fitness-based projection of long-term population growth can be made over any time-scale at which the environment is in a steady state and in which there are no irreversible changes. While there may be many fitness-based measures, it is crucial that these be constrained in two ways. First, as Game 2 makes clear, not just any arbitrarily identified reference environment will do; environments that admit irreversible changes must be barred. Second, restricting the set of permissible reference environments to those that are steady-state reversible does not ensure that different phenotypes will have different fitness.

STYMIED BY ECO-EVOLUTIONARY FEEDBACKS

A further problem faced by a general fitness value that could “average out/over” the specifics of particular environments involves the existence of eco-evolutionary feedbacks. With regard to our example of peppered moths, the problem would arise if we were to stipulate that a factory’s functional status (open versus closed) causally depends on the state of moths. However, examples of such feedback can be found across biological systems. Consider a pristine rainforest. The fitnesses of understory ferns with distinct character states can be compared under the assumption that the forest is at a steady state. The among-species comparisons (e.g., potential for competitive exclusion) of population ecology likewise rely on this assumption. A forest-clearing event would be an irreversible change in this context. Its occurrence would render projections based on the fitness estimates obtained in an unperturbed environment inaccurate, perhaps showing understory ferns “less fit” than some pioneer species. However, the presence of pioneer species changes the environment (e.g., by soil formation) and thus paves the way for the return of ferns during later stages of succession. Adequately comparing the fitnesses of a pioneer weed and an understory fern can only be done in a reference environment that encompasses periodic clearings and regrowth of the forest at a steady state. Examples such as this show how eco-evolutionary feedbacks further refine the constraints discussed in section 4: the inextricable coupling between changes in individuals and changes in their environments in biological populations shape what can be considered a set of inter-reachable environments (i.e. without irreversible changes). Overall, one cannot ignore feedbacks between the “ecological dynamics” that introduce new environments and the “evolutionary dynamics” that introduce new types[88]

Eco-evolutionary feedbacks constrain the choice of the reference environment because the set of accessible environments increases when new types appear. At a macroevolutionary scale, this fact, combined with resolution loss due to mutations, precludes a general definition of fitness (i.e., in an all-encompassing reference environment). To underscore this point, imagine comparing fitness values for mammals and cyanobacteria. Feedbacks require that this be done at a timescale over which even atmospheric composition changes (e.g., the cyanobacteria-driven oxygenation event 2.4 b.y.a.)[89] are reversible. As with the toy example presented in Section 5 (Figure 5b), the average value of fitness would be the same for any individual in the biosphere because, in the long run, it would not make any difference whether starting from a cyanobacterium or a mammal. The fitnesses of all organisms across the domain of life (past, present, and future) become indistinguishable.

Recent research shows that feedbacks between the evolutionary dynamics of individuals and the ecological dynamics of their environment are ubiquitous.[90–95] Recognizing as much obliges us to deal with both environmental and phenotypic reversibility when measuring fitness. The former type of reversibility might be seen as a welcome feature insofar as it permits a general measure of fitness to escape the trap posed by Game 2 (Figure 4e). But successfully accommodating environmental reversibility also introduces phenotypic inter-accessibility. This latter type of reversibility presents a seemingly insurmountable difficulty. For maintaining phenotypic accessibility in the long term simply reintroduces the predicament due to mutation discussed in Section 5 (Figure 5b), albeit this time in a somewhat less contrived and more immediate manner. In the limit, eco-evolutionary feedbacks guarantee that averaging over all possible environments necessarily leads to an average fitness value that is one and the same for all possible individuals.

ORGANISM-ENVIRONMENT INTERDEPENDENCIES: CONSEQUENCES FOR FITNESS MEASURES

The propensity interpretation of fitness posits that fitness measures the probabilistic capacity of an individual to produce offspring. This commitment provides a concept of fitness that can be invoked in causal explanations of evolutionary dynamics. Without it, fitness would be an explanatorily impotent, post hoc redescription of the living world (Section 2). However, the reproductive propensity of an individual is difficult to access, whether by means of statistical measurements on samples or via ecological knowledge of the system, and so fitness measures necessarily remain nonexhaustive summaries of this propensity (Section 3). The most salient difficulties for fitness on this interpretation arise from the fact that propensities are fundamentally relational properties of organism-environment pairings. In order to progress from descriptive statements in much delimited circumstances to explanations that enable prediction over multiple environments or individuals, reproductive outputs taken across different contexts must be adequately combined. This is often done by computing (potentially weighted) average reproductive outputs across multiple environments. Yet, as illustrated by the die-and-coins game presented in Figure 4, averaging reproductive outputs is not always adequate and may empty the value obtained of its predictive power. In particular, the strategy of averaging over all possible environments leads to erroneous results if there are *irreversible* environmental changes (Section 4). One way to alleviate this problem is to suppose that fitness measures only reflect long-term reproductive success, which occurs on a time scale over which all possible environments can be reached. However, this incurs a steep cost in resolution as the fitness values of similar, but different, individuals become equal (Section 5). These difficulties thwart any hope of computing a general scalar value of fitness that can distinguish selectively nonneutral differences between individuals whilst retaining its applicability as a predictor in any environment or evolutionary scenario. Indeed, feedback between the evolutionary dynamics of individuals and the ecological dynamics of their environment ensures that, in the limit, averaging over all possible environments would necessarily lead to an average fitness value for all possible individuals (Section 6).

CONCLUSION AND OUTLOOK: A CASE FOR ADAPTIVE DYNAMICS

Where do we go from here? None of what has been argued above dooms the propensity interpretation of fitness. Nor should it be taken as suggesting that general evolutionary mechanisms will forever remain obscure. Our intention is to convey no more or less than the following: if fitness is to retain the central conceptual role that it currently plays in evolutionary theory—as both reflecting historical success and enabling accurate prediction of future representation—then its measurement must be restricted to an intermediate number of inter-reachable environments and a timescale at which the blurring effect of mutation can be safely ignored. This proposal is not, of course, novel.[27] In practice, these constraints are already taken into account by the practitioners of adaptive dynamics. This family of methods distinguishes the ecological timescale, at which a unique measure of fitness ("invasion fitness") can be safely computed, from an evolutionary timescale that admits of prediction only via combining step-wise or "stringing together" the results of multiple ecological outcomes.[96–100] Adaptive dynamics, along with its preferred measure of fitness, is no silver bullet though. Just as classical irreversible thermodynamics is but one of the available theoretical frameworks for dealing with non-equilibrium thermodynamics,[101] so, too, is adaptive dynamics but one framework with the capacity to sidestep the difficulties noted in this paper. When the pivotal assumptions of adaptive dynamics do not hold, it may very well be the case that no scalar measure of fitness can accurately explain or predict the evolutionary dynamics. In such cases, one may have to resort to a more precise summary of the reproductive propensity, such as birth and death rates.[43,44,88,102]

ORCID

Pierrick Bourrat https://orcid.org/0000-0002-4465-6015

Notes

[1] The notion of propensity is contentious in philosophy of probability.[58] Some have tried to do away with it by proposing alternative interpretations of probabilities.[59–62] Several of these problems are inherited by the propensity interpretation of fitness.[63,64] Addressing such issues is beyond the scope of the present work.

[2] For the mathematically inclined reader, this game can be described by a discrete-time Markov chain with four states (blue heads, red heads, blue tails, red tails). The limit average score per round of a player (limit of Y_n for increasing n), is a linear combination of the sample average time spent in each state. The ergodic theorem (see Theorem 7.12 in Privault[80]) states that the sample average time spent in a given state converges toward the unique stationary distribution (and thus is independent of the initial conditions) if the chain is irreducible and positive recurrent (which implies that there are no irreversible changes in state).

[3] The Markov chain in such a case is not irreducible, so the ergodic theorem does not apply, and the long-term score does depend on the first toss.

ACKNOWLEDGMENTS

We owe thanks to Andrew Moore, the Theory and Method in Biosciences group at The University of Sydney and anonymous reviewers for feedback at various stages of this project. We would also like to thank Charles Pence for helpful comments and discussion.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

1. Abrams, M. (2012). Measured, modeled, and causal conceptions of fitness. *Front. Genet.*, *3*, 196. https://doi.org/10.3389/fgene.2012.00196
2. Coulson, T., Benton, T. G., Lundberg, P., Dall, S. R. X., Kendall, B. E., & Gaillard, J.-M. (2006). Estimating individual contributions to population growth: Evolutionary fitness in ecological time. *Proc. R. Soc. B. Biol. Sci.*, *273*(1586), 547–555. https://doi.org/10.1098/rspb.2005.3357
3. Dawkins, R. (1982). *The extended phenotype: The gene as the unit of selection*. Oxford University Press.
4. De Jong, G. (1994). The fitness of fitness concepts and the description of natural selection. *Q. Rev. Biol.*, *69*(1), 3–29. https://doi.org/10.1086/418431
5. Endler, J. A. (1986). *Natural selection in the wild*. (Vol. *21*). Princeton University Press.
6. Griffiths, J. I., Childs, D. Z., Bassar, R. D., Coulson, T., Reznick, D. N., & Rees, M. (2020). Individual differences determine the strength of ecological interactions. *Proc. Natl. Acad. Sci. USA*, *117*(29),17068–17073. https://eprints.whiterose.ac.uk/161577/
7. Hansen, T. F. (2017). On the definition and measurement of fitness in finite populations. *J. Theor. Biol.*, *419*, 36–43. https://doi.org/10.1016/j.jtbi.2016.12.024
8. Metz, J. A. J., Nisbet, R. M., & Geritz, S. A. H. (1992). How should we define 'fitness' for general ecological scenarios? *Trends Ecol. Evol.*, *7*(6), 198–202. https://doi.org/10.1016/0169-5347(92)90073-K
9. Michod, R. E. (1999). *Darwinian Dynamics*. Princeton: Princeton University Press.
10. Orr, H. A. (2009). Fitness and its role in evolutionary genetics. *Nat. Rev. Genet.*, *10*(8), 531–539. https://doi.org/10.1038/nrg2603
11. Tuljapurkar, S., Gaillard, J.-M., & Coulson, T. (2009). From stochastic environments to life histories and back. *Philos. Trans. R. Soc. B Biol. Sci.*, *364*(1523), 1499–1509. https://doi.org/10.1098/rstb.2009.0021
12. Tuljapurkar, S., Zuo, W., Coulson, T., Horvitz, C., & Gaillard, J.-M. (2020). Skewed distributions of lifetime reproductive success: Beyond mean and variance. *Ecol. Lett.*, *23*(4), 748–756. https://doi.org/10.1111/ele.13467
13. Van Valen, L. M. (1989). Three paradigms of evolution. *Evol. Theory*, *9*, 1–17.

14. Wagner, G. P. (2010). The measurement theory of fitness. *Evolution.*, *64*(5), 1358–1376. https://doi.org/10.1111/j.1558-5646.2009.00909.x
15. Rosenberg, A., & Bouchard, F. (2015). Fitness. In Edward N. Zalta (ed.), *The stanford encyclopedia of philosophy*. https://plato.stanford.edu/archives/spr2020/entries/fitness
16. Conner, J. K., & Hartl, D. L. (2004). *A Primer of Ecological Genetics*. Sunderland, Massachusetts: Sinauer Associates Incorporated.
17. Stearns, S. C. (1976). Life-history tactics: A review of the ideas. *Q. Rev. Biol.*, *51*(1), 3–47. https://doi.org/10.1086/409052
18. Bourrat, P. (2020). Natural selection and the reference grain problem. *Stud. Hist. Philos. Sci. Part A.*, *80*:1–8. https://doi.org/10.1016/j.shpsa.2019.03.003
19. Popper, K. R. (1974). Intellectual autobiography. *The Philosophy of Karl Popper*, *92*. https://ci.nii.ac.jp/naid/10004481309/
20. Scriven, M. (1959). Explanation and prediction in evolutionary theory: Satisfactory explanation of the past is possible even when prediction of the future is impossible. *Science*, *130*(3374), 477–482. https://doi.org/10.1126/science.130.3374.477
21. Fisher, R. A. (1930). *The genetical theory of natural selection*. Oxford: The Clarendon Press.
22. Haldane, J. B. (1932). *The causes of evolution*. Longmans: Green & Co. Limited.
23. Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, *6*, 97–159.
24. Brandon, R. N. (1978). Adaptation and evolutionary theory. *Stud. Hist. Philos. Sci. Part A.*, *9*(3), 181–206.
25. Brandon, R. N. (1990). *Adaptation and Environment*. Princeton, New Jersey: Princeton University Press.
26. Mills, S. K., & Beatty, J. H. (1979). The propensity interpretation of fitness. *Philos Sci.*, *46*(2), 263–286.
27. Pence, C. H., & Ramsey, G. (2013). A new foundation for the propensity interpretation of fitness. *Br J Philos Sci.*, *64*(4), 851–881. https://doi.org/10.1093/bjps/axs037
28. Ramsey, G. (2006). Block fitness. *Stud. Hist. Philos. Sci. Part C: Stud. Hist. Philos. Biol. Biomed. Sci.*, *37*(3), 484–498. https://doi.org/10.1016/j.shpsc.2006.06.009
29. Sober, E. (1984). Fact, fiction, and fitness: A reply to rosenberg. *J. Philos.*, *81*(7), 372–383. https://doi.org/10.2307/2026292
30. Millstein, R. L. (2016). Probability in biology: The case of fitness. In: A. Hájek & C. R. Hitchcock (Eds.), *The Oxford Handbook of Probability and Philosophy*. (pp. 601–622). Oxford: Oxford University Press.
31. Smart, J. J. C. (2014). *Philosophy and Scientific Realism*. Routledge.
32. Manser, A. R. (1965). The concept of evolution. *Philosophy*, *40*(151), 18–34.
33. Williams, M. B. (1970). Deducing the consequences of evolution: A mathematical model. *J. Theor. Biol.*, *29*(3), 343–385. https://doi.org/10.1016/0022-5193(70)90103-7
34. Rosenberg, A. (1982). On the propensity definition of fitness. *Philos. Sci.*, *49*(2), 268–273. https://www.journals.uchicago.edu/doi/abs/10.1086/289056?journalCode=phos
35. Rosenberg, A. (1983). Fitness. *J. Philos.*, *80*(8), 457–473. https://doi.org/10.2307/2026163
36. Rosenberg, A., & Williams, M. (1986). Fitness as primitive and propensity. *Philos. Sci.*, *53*(3), 412–418. https://doi.org/10.1086/289326
37. Brandon, Robert, & Beatty, John. (1984). The propensity interpretation of 'Fitness'–no interpretation is no substitute. *Philos. Sci.*, *51*(2), 342–347. https://www.journals.uchicago.edu/doi/abs/10.1086/289184?journalCode=phos
38. Thoday, B. J., M. (1953). Components of fitness. *Symp. Soc. Exp. Biol.*, (7), 96–113.
39. Gillespie, J. H. (1974). Nautural selection for within-generation variance in offspring number. *Genetics*, *76*(3), 601–606.
40. Gillespie, J. H. (1977). Natural selection for variances in offspring numbers: a new evolutionary principle. *Am. Nat.*, *111*(981), 1010–1014. https://doi.org/10.1086/283230
41. Gillespie, J. H. (1975). Natural selection for within-generation variance in offspring number II. Discrete haploid models. *Genetics*, *81*(2), 403–413.
42. Levins, R. (1968). *Evolution in changing environments: Some theoretical explorations*. Princeton, NJ: Princeton University Press.
43. Lewontin, R. C., & Cohen, D. (1969). On population growth in a randomly varying environment. *Proc. Natl. Acad. Sci. U S A*, *62*(4), 1056–1060.
44. Beatty, J., & Finsen, S. (1989). Rethinking the propensity interpretation: A peek inside pandora's Box1. In: M. Ruse (Ed.), *What the Philosophy of Biology Is: Essays dedicated to David Hull*. (pp. 17–30). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-009-1169-7_2
45. Sober, E. (2001). The two faces of fitness. In: R. Singh Costas B. Krimbas, Diane B. Paul, & John Beatty (Ed.), *Thinking about evolution: Historical, philosophical, and political perspectives (Festchrifft for Richard C. Lewontin Vol. 2)*. (pp. 309–321). Cambridge: Cambridge University Press.
46. Matthen, M., & Ariew, A. (2002). Two ways of thinking about fitness and natural selection. *J. Philos.*, *99*(2), 55–83.
47. Walsh, D. M., Lewens, T., & Ariew, A. (2002). The trials of life: Natural selection and random drift. *Philos. Sci.*, *69*(3), 429–446.
48. Ariew, A., & Lewontin, R. C. (2004). The confusions of fitness. *Br. J. Philos. Sci.*, *55*(2), 347–363. https://doi.org/10.1093/bjps/55.2.347
49. Ariew, A., & Ernst, Z. (2009). What fitness can't be. *Erkenntnis*, *71*(3), 289.
50. Walsh, D. M. (2010). Not a sure thing: Fitness, probability, and causation. *Philos. Sci.*, *77*(2), 147–171. https://doi.org/10.1086/651320
51. Walsh, D. M. (2007). The pomp of superfluous causes: The interpretation of evolutionary theory. *Philos. Sci.*, *74*(3), 281–303.
52. Abrams, M. (2007). Fitness and propensity's annulment? *Biol. Philos.*, *22*(1), 115–130.
53. Abrams, M. (2009). The unity of fitness. *Philos. Sci.*, *76*(5), 750–761.
54. Abrams, M. (2009). Fitness "kinematics": Biological function, altruism, and organism–environment development. *Biol. Philos.*, *24*(4), 487–504. https://doi.org/10.1007/s10539-009-9153-2
55. Otsuka, J., Turner, T., Allen, C., & Lloyd, E. A. (2011). Why the causal view of fitness survives. *Philos. Sci.*, *78*(2), 209–224. https://doi.org/10.1086/659219
56. Millstein, R. L. (2006). Natural selection as a population-level causal process. *Br. J. Philos. Sci.*, *57*(4), 627–653.
57. Tuljapurkar, S. (1989). An uncertain life: Demography in random environments. *Theor. Popul. Biol.*, *35*(3), 227–294. https://doi.org/10.1016/0040-5809(89)90001-4
58. Hájek, A. (2012). Interpretations of probability. In: E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*. (Winter 2012.). http://plato.stanford.edu/archives/win2012/entries/probability-interpret/
59. Abrams, M. (2012). Mechanistic probability. *Synthese*, *187*(2), 343–375. https://doi.org/10.1007/s11229-010-9830-3
60. Lyon, A. (2011). Deterministic probability: Neither chance nor credence. *Synthese*, *182*, 413–432.
61. Rosenthal, J. (2010). The natural-range conception of probability. In: G. Ernst & A. Hüttemann (Eds.), *Time, chance, and reduction: Philosophical aspects of statistical mechanics*. (pp. 71–90). Cambridge, UK; New York: Cambridge University Press.
62. Strevens, M. (2011). Probability out of determinism. In: C. Beisbart & S. Hartman (Eds.), *Probabilities in physics*. (pp. 339–364). Oxford: Oxford University Press.
63. Bourrat, P. (2017). Explaining drift from a deterministic setting. *Biol. Theory.*, *12*(1), 27–38.
64. Godfrey-Smith, P. (2009). *Darwinian populations and natural selection*. Oxford; New York: Oxford University Press.

65. Beatty, J. H. (1984). Chance and natural selection. *Philos. Sci.*, *51*(2), 183–211.
66. Bourrat, P. (2018). Natural selection and drift as individual-level causes of evolution. *Acta Biotheor.*, *66*(3), 159–176. https://doi.org/10.1007/s10441-018-9331-1
67. Hodge, M. J. S. (1987). Natural selection as a causal, empirical, and probabilistictheory. In *The probabilistic revolution*. (pp. 233–270). Cambridge MA: MIT Press.
68. Plutynski, A. (2007). Drift: A historical and conceptual overview. *Biol. Theory*, *2*, 156–167. https://doi.org/10.1162/biot.2007.2.2.156
69. Ramsey, G. (2013). Driftability. *Synthese*, *190*(17), 3909–3928. https://doi.org/10.1007/s11229-012-0232-6
70. Popper, K. R. (1959). The propensity interpretation of probability. *Br. J. Philos. Sci.*, *10*(37), 25–42.
71. Strevens, M. (2013). *Tychomancy inferring probability from causal structure*. Cambridge, MA: Harvard University.
72. Airoldi, E. M., Huttenhower, C., Gresham, D., Lu, C., Caudy, A. A., Dunham, M. J., … Troyanskaya, O. G. (2009). Predicting cellular growth from gene expression signatures. *PLoS Comput. Biol.*, *5*(1). https://doi.org/10.1371/journal.pcbi.1000257
73. Quinn, G. P., & Keough, M. J. (2002). *Experimental design and data analysis for biologists*. Cambridge, UK ; New York: Cambridge University Press.
74. Rohlf, F. J., & Sokal, R. R. (1995). *Biometry: The principles and practice of statistics in biological research*. New York: W.H. Freeman and Company.
75. Keyfitz, N. (1972). On future population. *J. Am. Statist. Assoc.*, *67*(338), 347–363. https://doi.org/10.1080/01621459.1972.10482386
76. Bourrat, P. (2015). Levels, time and fitness in evolutionary transitions in individuality. *Philos. Theory Biol.*, *7*(1). https://doi.org/10.3998/ptb.6959004.0007.001
77. Bourrat, P. (2015). Levels of selection are artefacts of different fitness temporal measures. *Ratio*, *28*(1), 40–50.
78. Kettlewell, H. B. D. (1955). Selection experiments on industrial melanism in the Lepidoptera. *Heredity*, *9*, 323–342.
79. Majerus, M. E. N. (1998). *Melanism: Evolution in action*. Oxford University Press.
80. Privault, N. (2018). *Understanding markov chains: Examples and applications*. (2nd ed.). Springer, Singapore. https://doi.org/10.1007/978-981-13-0659-4
81. Millstein, R. L. (2017). Genetic drift. In: E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*. (Fall 2017.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2017/entries/genetic-drift/
82. Millstein, R. L. (2002). Are random drift and natural selection conceptually distinct? *Biol. Philos*, *17*, 33–53. https://doi.org/10.1023/a:1012990800358
83. Sober, E. (2020). Fitness and the twins. *Philos, Theory, Pract. Biol.*, *12*(1). https://doi.org/10.3998/ptpbio.16039257.0012.001
84. Masel, J. (2012). Rethinking Hardy-Weinberg and genetic drift in undergraduate biology. *BioEssays*, *34*(8), 701–710. https://doi.org/10.1002/bies.201100178
85. Bourrat, P. (2015). Distinguishing natural selection from other evolutionary processes in the evolution of altruism. *Biol. Theory*, *10*(4), 311–321.
86. Pence, C. H., & Ramsey, G. (2015). Is organismic fitness at the basis of evolutionary theory? *Philos. Sci.*, *82*(5), 1081–1091. https://doi.org/10.1086/683442
87. Abrams, M. (2009). What determines biological fitness? The problem of the reference environment. *Synthese*, *166*(1), 21–40.
88. Sober, E. (2006). The two faces of fitness. *Concept Issues Evol. Biol.*, 25–38.
89. Lyons, T. W., Reinhard, C. T., & Planavsky, N. J. (2014). The rise of oxygen in Earth's early ocean and atmosphere. *Nature*, *506*(7488), 307–315. https://doi.org/10.1038/nature13068
90. Coulson, T., Potter, T., & Felmy, A. (2019). Fitness functions, genetic and non-genetic inheritance, and why ecological dynamics and evolution are inevitably linked. *bioRxiv*, 762658. https://doi.org/10.1101/762658
91. Ellner, S. P., Geber, M. A., & Hairston, N. G. (2011). Does rapid evolution matter? Measuring the rate of contemporary evolution and its impacts on ecological dynamics. *Ecol. Lett.*, *14*(6), 603–614. https://doi.org/10.1111/j.1461-0248.2011.01616.x
92. Kokko, H., & López-Sepulcre, A. (2007). The ecogenetic link between demography and evolution: Can we bridge the gap between theory and data? *Ecol. Lett.*, *10*(9), 773–782. https://doi.org/10.1111/j.1461-0248.2007.01086.x
93. Reznick, D. N., & Ghalambor, C. K. (2001). The population ecology of contemporary adaptations: What empirical studies reveal about the conditions that promote adaptive evolution. In: A. P. Hendry & M. T. Kinnison (Eds.), *Microevolution Rate, Pattern, Process*. (pp. 183–198). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-010-0585-2_12
94. Schoener, T. W. (2011). The newest synthesis: Understanding the interplay of evolutionary and ecological dynamics. *Science*, *331*(6016), 426–429. https://doi.org/10.1126/science.1193954
95. Travis, J., Reznick, D., Bassar, R. D., López-Sepulcre, A., Ferriere, R., & Coulson, T. (2014). Chapter one—do eco-evo feedbacks help us understand nature? Answers from studies of the Trinidadian Guppy. In: J. Moya-Laraño, J. Rowntree, & G. Woodward (Eds.), *Advances in Ecological Research*. (Vol. *50*, pp. 1–40). Academic Press. https://doi.org/10.1016/B978-0-12-801374-8.00001-3
96. Dieckmann, U., & Law, R. (1996). The dynamical theory of coevolution: A derivation from stochastic ecological processes. *J. Math. Biol.*, *34*(5), 579–612. https://doi.org/10.1007/BF02409751
97. Geritz, S. A. H., Kisdi, E., Meszéna, G., & Metz, J. a. J. (1998). Evolutionarily singular strategies and the adaptive growth and branching of the evolutionary tree. *Evol. Ecol.*, *12*(1), 35–57. https://doi.org/10.1023/A:1006554906681
98. Waxman, D., & Gavrilets, S. (2005). 20 questions on adaptive dynamics: Target review of adaptive dynamics. *J. Evol. Biol.*, *18*(5), 1139–1154. https://doi.org/10.1111/j.1420-9101.2005.00948.x
99. Brännström, Å., Johansson, J., & von Festenberg, N. (2013). The Hitchhiker's guide to adaptive dynamics. *Games*, *4*(3), 304–328. https://doi.org/10.3390/g4030304
100. Lion, S. (2017). Theoretical approaches in evolutionary ecology: Environmental feedback as a unifying perspective. *Am. Nat.*, *191*(1), 21–44. https://doi.org/10.1086/694865
101. Lebon, G., Jou, D., & Casas-Vázquez, J. (2008). *Understanding non-equilibrium thermodynamics: Foundations, applications, frontiers*. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-74252-4
102. Yoshimura, J., & Clark, C. W. (2012). *Adaptation in stochastic environments*. (Vol. 98). Springer Science & Business Media.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Doulcier, G., Takacs, P., & Bourrat, P. (2021). Taming fitness: Organism-environment interdependencies preclude long-term fitness forecasting. *Bioessays*, 43, e2000157. https://doi.org/10.1002/bies.202000157