Supplementary materials

Relating to the paper: Henshaw J.M., M. B. Morrissey, and A. G. Jones. Quantifying the causal pathways contributing to natural selection. Evolution.

Here we provide formal proofs for several results in the main text.

Causal derivatives using Pearl's do-calculus

Here we provide an alternative definition of causal derivatives $\frac{\delta w}{\delta Z}$ using the do-calculus, which is a component of Pearl's theory of causal inference (Pearl 2009, 2018). An *intervention* on a trait Z can be expressed as an operation of the form $\mathrm{do}(Z = z)$, where the value of Z is set to z without regard for how Z would ordinarily be determined. This is equivalent to replacing the function f_Z with the constant $Z = z$ in the causal model. We write $Y_{\mathrm{do}(Z=z)}$ for the random variable that would be obtained by intervening to set Z equal to z and then observing Y. In particular, suppose we change a trait Z by a small quantity ε using the intervention $\mathrm{do}(Z = Z + \varepsilon)$. The absolute change in fitness due to this intervention is $w_{\mathrm{do}(Z=Z+\varepsilon)} - w$. The *rate* at which fitness changes with the respect to the change in Z is $\frac{1}{\varepsilon}\left(w_{\mathrm{do}(Z=Z+\varepsilon)} - w\right)$. Now, the causal derivative $\frac{\delta w}{\delta Z}$ is the rate at which fitness changes under an infinitesimal intervention on Z. Hence, an alternative way to express the causal derivative is:

$$\frac{\delta w}{\delta Z} = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}\left(w_{\mathrm{do}(Z=Z+\varepsilon)} - w\right) \tag{S1}$$

Some purists may object to the notation $\mathrm{do}(Z = Z + \varepsilon)$, where the intervention appears to depend on the realised value of Z. To render this notation harmless, we can instead modify the causal graph $\mathcal{G}$ by introducing an instrumental variable I_Z that affects only Z, and replacing the function f_Z by $\widehat{f_Z}(\cdot, I_Z) = f_Z(\cdot) + I_Z$ (Pearl 2009). In other words, I_Z allows one to add or subtract from the value of Z. The intervention $\mathrm{do}(Z = Z + \varepsilon)$ is then equivalent to $\mathrm{do}(I_Z = \varepsilon)$.

Proof that $S_{\mathrm{causal},Z} = \sigma_Z^2 \eta_Z$

Here we show that for normally distributed traits Z, the component of the selection differential that arises from the causal effects of Z on fitness is given by $\sigma_Z^2 \eta_Z$. First, we note that by Stein's lemma (Stein 1981; Liu 1994; de Villemereuil et al. 2016; Walsh and Morrissey 2019), the selection differential on Z can be written as:

$$S_Z = \sigma_Z^2 \mathbb{E}\left(\frac{d}{dZ}\mathbb{E}(w|Z)\right) \tag{S2}$$

By the law of total expectation, we have:

$$\mathbb{E}(w|Z) = \int \mathbb{E}(w|Z, \mathrm{pa}(Z), U_Z)\; d\mathbb{P}(\mathrm{pa}(Z), U_Z|Z) \tag{S3}$$

Substituting this expression into equation (S2), we can differentiate under the integral sign and apply the product rule of calculus to obtain:

$$S_Z = \sigma_Z^2 \mathbb{E}\left(\int \frac{\partial}{\partial Z}\mathbb{E}(w|Z, \mathrm{pa}(Z), U_Z)\ \ d\mathbb{P}(\mathrm{pa}(Z), U_Z|Z) + \int \mathbb{E}(w|Z, \mathrm{pa}(Z), U_Z)\frac{d}{dZ}d\mathbb{P}(\mathrm{pa}(Z), U_Z|Z)\right) \tag{S4}$$

The first term tells us how small changes in Z affect fitness, controlling for the causal determinants of Z. The second term tells us how correlated changes in the causal determinants of Z are associated with changes in fitness, while holding the trait Z itself constant. Equation (S4) thus decomposes the selection differential into causal and spurious components. The first term simplifies to:

$$S_{\mathrm{causal},Z} = \sigma_Z^2 \mathbb{E}\left(\frac{\partial}{\partial Z}\mathbb{E}(w|Z, \mathrm{pa}(Z), U_Z)\right) \tag{S5}$$

This term represents how small changes in Z affect fitness, controlling for the causal determinants of Z. Since $\{\mathrm{pa}(Z), U_Z\}$ is a backdoor set for Z, we can re-write the right-hand side as $\sigma_Z^2\eta_Z$.

Proof of the first-edge criterion for the identifiability of path-specific selection gradients
Suppose $\mathcal{H}$ is a subgraph of $\mathcal{G}$ consisting of the union of any number of directed paths from Z to w. We assume that the background variables $\boldsymbol{U_Y}$ of endogenous traits are mutually independent and that $\boldsymbol{U_Y}$ is independent of the exogenous traits $\boldsymbol{X}$. Avin et al. (2005) proved the following elegant theorem:

(Recanting witness criterion): Suppose $\mathcal{G}$ contains a *recanting witness*, defined as a variable R such that (i) there is a path in $\mathcal{H}$ that passes from Z through R to w, and (ii) there is a path from R to w that is in $\mathcal{G}$ but not in $\mathcal{H}$. Then the $\mathcal{H}$-specific selection gradient of Z on w is *not* identifiable. Conversely, if no such variable R exists, then the $\mathcal{H}$-specific selection gradient is identifiable.

The odd name of this criterion refers to the following metaphor: We can think of the variable R as a ‘witness’ to Z, in that it sees the causal effect of Z and reports it on to downstream variables like fitness (more prosaically, R mediates at least some of the effects of Z on w). Suppose that now that R reports to w along two separate causal pathways. To tease apart these pathways, we must imagine R reporting one thing along the first pathway and a different thing along the second (e.g. the effects of a manipulated or an unmanipulated Z, respectively). The variable R is a ‘recanting witness’ in this scenario because it ‘changes its mind’ about what to report. Path-specific effects that can only be conceptualised via such recanting witnesses cannot be estimated from observational data (i.e. they are unidentifiable).

The original proof of the recanting witness criterion applies strictly only to effects of the form $\mathbb{E}_{\boldsymbol{U}}\left(\frac{\delta w}{\delta Z}\Big|_{\mathcal{H}}\right)$ in *Markovian* models, where all background variables $\boldsymbol{U}$ are mutually independent. However, we can accommodate covariation among the exogenous traits $\boldsymbol{X}$ by first calculating $\mathbb{E}_{\boldsymbol{U_Y}}\left(\frac{\delta w}{\delta Z}\Big|_{\mathcal{H}}\right)$ for fixed values of the exogenous traits $\boldsymbol{X}$ and then integrating over these values to obtain the $\mathcal{H}$-specific selection gradient:

$$\eta_Z|_{\mathcal{H}} = \mathbb{E}_{\boldsymbol{X}}\left(\mathbb{E}_{\boldsymbol{U_Y}}\left(\frac{\delta w}{\delta Z}\bigg|_{\mathcal{H}}\right)\right) \tag{S6}$$

Note that this requires the independence of $\boldsymbol{U_Y}$ and $\boldsymbol{X}$. The criterion is also robust to some forms of non-independence among the background variables $\boldsymbol{U_Y}$, but we do not consider this relaxation here.

The first-edge criterion in the main text is a straightforward corollary of the recanting witness criterion. First, suppose C is a child of Z and $\mathcal{H}(C)$ is the subgraph of $\mathcal{G}$ consisting of all directed paths from Z to w whose first edge is $Z \to C$. If R is any variable in $\mathcal{H}(C)$ other than Z or w, then by construction there must exist a directed path P from Z to R whose first edge is $Z \to C$. For any directed path P^* from R to w, we can then construct a new directed path $Z \to_P R \to_{P^*} w$, which has $Z \to C$ as its first edge. Hence, P^* must lie in $\mathcal{H}(C)$. Therefore, R is not a recanting witness, and the C-specific selection gradient is identifiable. It follows easily that if $\mathcal{H} = \bigcup_{C \in \boldsymbol{C}} \mathcal{H}(C)$, where $\boldsymbol{C}$ is a subset of the children Z, then the $\mathcal{H}$-specific selection gradient is also identifiable. Conversely, suppose $\mathcal{H}$ is a subgraph of $\mathcal{G}$ that cannot be expressed as such a union. Then there must exist some child C of Z such that at least one path with first edge $Z \to C$ lies in $\mathcal{H}$, and at least one such path does not fully lie in $\mathcal{H}$. These two paths satisfy conditions (i) and (ii) of the recanting witness criterion and so the $\mathcal{H}$-specific selection gradient is not identifiable.

Estimation of path-specific selection gradients in Markovian models

Here we prove equations (15) and (16) of the main text, which give explicit formulas for the extended and path-specific selection gradients. We assume that the background variables $\boldsymbol{U_Y}$ of the endogenous traits are both mutually independent and independent of $\boldsymbol{X}$. This implies the Markovian property given by equation (13) of the main text. We proceed from the definition of causal derivatives given in equation (S1). By integrating under the limit, the extended selection gradient on Z can then be expressed as:

$$\eta_Z = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \mathbb{E}\big(w_{\mathrm{do}(Z=Z+\varepsilon)} - w\big) \tag{S7}$$

We note that if the original model $\mathcal{M}$ obeys the Markovian property, then so does the extended model $\widehat{\mathcal{M}}$ in which Z is given a new parent I_Z and the function f_Z is replaced with $\widehat{f_Z}(\cdot, I_Z) = f_Z(\cdot) + I_Z$ (see above). We can then write:

$$\begin{aligned} &\mathbb{E}\big(w_{\mathrm{do}(Z=Z+\varepsilon)}\big) \\ &= \int w \;\; d\mathbb{P}(\boldsymbol{X}) \left(\prod_{Y \in \boldsymbol{Y} \setminus \mathrm{ch(Z)}} d\mathbb{P}\big(Y \big| \mathrm{pa}(Y)\big) \right) \left(\prod_{C \in \mathrm{ch(Z)}} d\mathbb{P}\big(C \big| \mathrm{pa}(C)\big)\big|_{Z=Z+\varepsilon} \right) \end{aligned} \tag{S8}$$

and similarly for $\mathbb{E}(w)$. Substituting these expressions into equation (S7) and evaluating the limit gives us equation (15) in the main text.

We next formulate a 'twin network' model to quantify the path-specific selection gradient on Z via its child C. First, let us modify the original graph $\mathcal{G}$ by (i) removing Z along with its associated edges and replacing it with two new variables Z_C and Z_* (ii) adding an edge from Z_C to C, and (iii) adding edges from Z_* to all other former children of Z, excluding C (see

Figure 1b in the main text). We modify the original model $\mathcal{M}$ correspondingly, by (iv) replacing the function f_C by a new function f_C^* that takes Z_C instead of Z as an argument, but is otherwise identical, and similarly (v) for all other former children Y of Z, replacing the function f_Y with a new function f_Y^* that takes Z_* as an argument instead of Z. We denote the modified graph $\mathcal{G}^*$ and the modified model $\mathcal{M}^*$. Essentially, the new model allows Z to take one value for the purposes of influencing its child C, but another value for influencing all other children.

The C-specific selection gradient on Z in the original model $\mathcal{M}$ can be found by first calculating the extended selection gradient on Z_C in the modified model $\mathcal{M}^*$ and then setting $Z_C = Z_* = Z$. In other words:

$$\eta_Z|_C = \eta_{Z_C}\big|_{Z_C=Z_*=Z} \tag{S9}$$

Evaluating the right-hand side using equation (15) from the main text then yields equation (16).

Proof that $\boldsymbol{\beta_A} = \boldsymbol{\eta_U}$

Suppose that (i) $f_{U_Z}(A_Z, R_Z) = A_Z + R_Z$ for all traits Z, (ii) $\boldsymbol{A}$ and $\boldsymbol{R}$ are independent, and (iii) w is independent of $\boldsymbol{A} \cup \boldsymbol{R}$ given $\boldsymbol{U}$ (i.e. w is *d*-separated from $\boldsymbol{A} \cup \boldsymbol{R}$ by $\boldsymbol{U}$: Geyer and Shaw 2008; Pearl 2009; Queller 2017). We will show that the average partial selection gradients on $\boldsymbol{A}$ equal the extended selection gradients: $\boldsymbol{\beta_A} = \boldsymbol{\eta_U}$. First, since $\boldsymbol{A}$ is independent of $\boldsymbol{R}$, the law of total expectation gives us:

$$\mathbb{E}(w|\boldsymbol{A}) = \mathbb{E}_{\boldsymbol{R}|\boldsymbol{A}}\mathbb{E}(w|\boldsymbol{A},\boldsymbol{R}) = \mathbb{E}_{\boldsymbol{R}}\mathbb{E}(w|\boldsymbol{A},\boldsymbol{R}) \tag{S10}$$

Substituting this quantity into the definition of $\boldsymbol{\beta_A}$ yields:

$$\boldsymbol{\beta_A} = \mathbb{E}\big(\nabla\mathbb{E}(w|\boldsymbol{A})\big) = \mathbb{E}\big(\nabla_{\boldsymbol{A}}\mathbb{E}_{\boldsymbol{R}}\mathbb{E}(w|\boldsymbol{A},\boldsymbol{R})\big) \tag{S11}$$

Using the independence of $\boldsymbol{A}$ from $\boldsymbol{R}$ once more, we can then pass the gradient $\nabla_{\boldsymbol{A}}$ under the expectation $\mathbb{E}_{\boldsymbol{R}}$ to obtain:

$$\boldsymbol{\beta_A} = \mathbb{E}\big(\nabla_{\boldsymbol{A}}\mathbb{E}(w|\boldsymbol{A},\boldsymbol{R})\big) \tag{S12}$$

Lastly, since $\boldsymbol{U}$ is fully determined by $\boldsymbol{A}$ and $\boldsymbol{R}$, and w is independent of $\boldsymbol{A} \cup \boldsymbol{R}$ given $\boldsymbol{U}$, we have $\mathbb{E}(w|\boldsymbol{A},\boldsymbol{R}) = \mathbb{E}(w|\boldsymbol{U})$. Using the assumption that effects are additive (assumption (i) above) and the chain rule of calculus, we thus obtain:

$$\boldsymbol{\beta_A} = \mathbb{E}\big(\nabla\mathbb{E}(w|\boldsymbol{U})\big) = \boldsymbol{\eta_U} \tag{S13}$$

Similar results for the equivalence of linear regression selection gradients on traits and breeding values have a long history in the literature (see Rausher 1992; Morrissey et al. 2010; Walsh and Lynch 2018, Ch.20).