Reading the Receipt, Not the Person: An Anti-Physiognomy Framing for Developmental Trace Inference

Michael Abdo
Independent Researcher · Correspondence: Michael@michaelabdo.com

Perspective.

Abstract

Physiognomy, the claim that a person's face causes or directly displays their character, is false, and modern machine-learning revivals of it have become a reputational and ethical landmine. Yet a structurally different claim hides in the same neighborhood: a single prenatal event (the intrauterine hormonal milieu acting across critical developmental windows) can shape both physical structure and neural architecture so that observable form becomes a correlated trace of a developmental signal rather than a cause of mind. This Perspective draws the bright line between the two. Physiognomy asserts body-to-mind causation and uniform legibility of character; the common-cause claim asserts a shared upstream cause, predicts a graded ordering of readability, and yields a probabilistic prior to be updated, never a verdict. The discipline separating the two reduces to three commitments: (1) pre-registration of the directional prediction, (2) control for demographic confounds (sex above all), and (3) honest reporting of one's own nulls. I use our own pre-registered test as the worked example: we predicted a facial readability hierarchy, registered it, tested it on held-out identities and it failed. The only signal whose bootstrap confidence interval excluded zero was the sex-modality axis, which reduces most plausibly to a sex-correlated demographic shortcut, precisely the failure mode physiognomy refuses to look at. A predicted-near-null letter axis also out-separated the predicted-strong function axis, inverting the ordering and sharpening the refutation. Reporting that null under pre-registration is itself the evidence that the line between the two claims is real.

---

1. The physiognomy landmine, and why it keeps detonating

Physiognomy is the old idea that the face reveals the character. A noble brow signals a noble mind. The criminal "type" can be read off the cheekbones. Few bad ideas in the history of human inquiry have proven so durable, and this one has a body count: it supplied a veneer of measurement to phrenology, to Lombroso's "born criminal," and to the racial sciences of the twentieth century. The reason it is pseudoscience is not that faces carry no information. They obviously carry some. The problem is that the *claim* it makes is the wrong shape. Physiognomy asserts that the body **causes** or **is** the mind. It claims that "character" is **uniformly legible** from the surface. It posits that any competent reader, looking at any face, can recover a verdict about the person behind it.

This claim resurfaces in every generation, as each new measurement technology provides a fresh way to launder the same inference, and the current vehicle is machine learning. The canonical modern episode is Wang and Kosinski's 2018 report that deep neural networks could distinguish sexual orientation from facial images at AUCs exceeding human-judge performance. They reported AUCs (area under the ROC curve, not classification accuracy) of 0.81 for men and 0.71 for women from a single image, rising to 0.91 for men given five images. This work was published in a flagship psychology journal (Wang & Kosinski, 2018). That the original AUCs were widely re-reported as

accuracy percentages is itself part of how the result came to be over-read. The result detonated immediately. Within months, Agüera y Arcas, Mitchell, and Todorov published a careful rebuttal showing that the apparent signal was almost entirely counterfeit. The classifier was keying on grooming, head pose, facial hair, makeup, eyeshadow, even the camera angle and the average brightness of self-presentation, not on any fixed facial morphology (Agüera y Arcas, Mitchell & Todorov, 2018). Gay and straight men in the dataset differed in how often they wore glasses. Not optics. Self-presentation, a choice about how they liked to look. The network had learned a culture, photographed. It had not read a body.

That episode is the landmine. It is worth being precise about why it is dangerous, because the danger is not only reputational. There are three distinct failure modes braided together:

1. **The confound mistaken for the trait.** What looks like a face→trait effect is really a face→demographic→trait chain, or a face→self-presentation→label chain. The model reads the proxy and reports the target.
2. **The uniform-legibility error.** Even where a population-level statistical association is real, physiognomy converts it into a claim about *every individual* - a verdict, applied one face at a time.
3. **The use case.** The inference is offered as a tool for screening, ranking, hiring, policing, or gating. At that point a noisy population correlation becomes a machine for inflicting individualized harm, and the harm lands hardest on the groups the confounds track.

Any research program that touches faces and dispositions inherits all three liabilities by default. The burden is on the researcher to show, affirmatively, that they have disarmed each one. Most cannot. That is exactly why the field has earned its reputation. This Perspective asks a narrower and more constructive question. Is there a claim in this space that is structurally different from physiognomy, not merely more careful but a different kind of statement altogether? And if such a claim exists, what discipline would let us tell the difference?

2. The common-cause alternative: trace, not cause

This claim is not new; it predates the machine-learning era. It does not assert that the face reveals the mind. Rather, it suggests that the face and the mind were built, in part, by the same event, and so they carry correlated traces of it.

The mechanism is the organizational hypothesis of developmental neuroendocrinology, established in animals by Phoenix, Goy, Gerall, and Young (1959) and extended to humans across decades of work: hormones present during critical fetal windows exert permanent, organizing effects on tissue and they organize multiple tissues at once. The intrauterine hormonal milieu (testosterone, cortisol, thyroid, estradiol, others, each varying in dose and in the timing of exposure) simultaneously shapes physical structure (including the bones of the face and the vocal apparatus) and wires neural architecture. One signal, two outputs: a body and a brain, made together, from the same developmental input.

This is not invention. The single-axis version is accepted science. Prenatal androgen excess in congenital adrenal hyperplasia virilizes anatomy and shifts behavior in the same individuals. It serves as the textbook natural experiment of one gradient and two correlated outputs (Berenbaum & Beltz, 2011). The most directly relevant result comes from Whitehouse and colleagues. They measured testosterone in umbilical cord blood at birth. Then in the same people 21-24 years later,

they recovered a medium-to-strong association between that prenatal hormone level and the sexual dimorphism of the adult three-dimensional face (Whitehouse et al., 2015). A hormone measured at birth left a readable footprint on the face two decades on. That is the existence proof that a face can be a trace of a prenatal developmental signal.

The conceptual move that separates this from physiognomy is causal and it is sharp. Write the developmental input as a vector h of hormone exposures. The model posits two coupled maps from the same argument: a morphological map M(h) leading to adult facial and vocal form and a neural map N(h) leading to dispositional architecture. Because M and N share h their outputs correlate with no causal arrow between body and mind. The cheekbone does not cause the temperament. Personality is not read off the face. Both are reads of h. The face is the receipt of a transaction you cannot observe directly, not the transaction itself.

It is a textbook example of a common-cause structure. And that changes everything downstream.

Trace, not cause. Physiognomy requires body-to-mind causation. Once you control for the actual upstream cause, its claim should evaporate. The common-cause claim expects the correlation to be non-causal between body and mind. It locates the causation entirely upstream, in development. This is not a semantic dodge. It makes different predictions, which appear below. It also forbids different things.

Ordering, not verdict. Physiognomy claims uniform legibility: every face yields a character. The common-cause model claims something far weaker and far stranger. It suggests that readability should be graded, and graded in a specific, pre-committable order. If the face is a trace of h, then trait-axes wired by loud hormonal gradients, the sexually dimorphic axes organized by large androgen differences, should be more face-readable than axes with diffuse or late hormonal input. Those latter axes should approach chance. The model predicts a hierarchy of readability that tracks an independently-known measure of hormonal organization. This is a prediction physiognomy never made and could not make. Confounds like image quality, demographics, and grooming do not respect that ordering. A counterfeit signal has no reason to arrange itself in the order prenatal endocrinology predicts.

And even a confirmed ordering buys you less than it sounds. A prior, not a judgment, is all it would license about any individual. It is the first word, never the last. A prior sets the odds; the person plays them. If you refuse to update a prior, it is no longer a prior but a prejudice. The model is a statement about populations and mechanism, not a license to judge a person.

The distinction isn't simply that we are more careful than physiognomists. It is this: physiognomy asserts a causal arrow that does not exist and a verdict it cannot support. Common-cause trace inference, by contrast, asserts a shared upstream cause, predicts a graded ordering, and yields an updatable prior. Two claims, different shapes, different obligations.

3. Three disciplines that separate science from physiognomy

A different shape of claim is necessary but not sufficient. A researcher could make a perfectly well-formed common-cause claim and still launder a confound, still over-read a fragile effect, still quietly bury the runs that failed. What converts the structural distinction into a scientific one is method. Three disciplines do the work. Each maps directly onto one of the three physiognomy failure modes from Section 1.

3.1 Pre-registration - against the uniform-legibility error and the garden of forking paths

The physiognomist's signature move is to find a signal and declare it the character. With high-dimensional facial data, which includes dozens of morphometric components, multiple trait axes, and a choice of which principal component to report, a researcher who looks first and theorizes after can almost always recover something that looks like a hierarchy. The defense is to commit the directional prediction, the estimator, the covariates, and the decision rule to a timestamped record before touching the held-out data, and then to run exactly that. Pre-registration does not make a result true. It makes a result honest. It forecloses the post hoc rescue, the silent switch from the component that failed to the one that worked, the redefinition of confirmation after the numbers are in. For a face-based claim, where the prior probability of self-deception is enormous, pre-registration is not best practice. It is the price of admission.

3.2 Demographic-confound control - against the confound mistaken for the trait

Demographics are the adversary, and sex is the chief adversary. Faces encode biological sex strongly, and almost any trait that correlates with sex in the population will appear "readable from the face" through that single channel, reproducing the Wang-Kosinski failure exactly. The discipline has three layers. The first is to read sensitive demographics from outside the image, such as biological sex from a database record rather than from a face model, so the confound cannot leak through the same channel as the signal. Then residualize: remove the sex direction from the facial representation, re-estimate, and report raw and adjusted side by side. The third layer is the most demanding, and it is where the whole distinction lives. Replicate within each demographic stratum, within each sex, within each ethnicity, because an effect that recurs with the same sign inside a single-sex subgroup cannot be the sex confound wearing a disguise. The discipline of partialling is not optional polish. It is the operationalization of the entire trace-versus-confound distinction. An effect that vanishes under sex adjustment was never a trait read. It was sex, read.

3.3 Reporting your own nulls - against the use case, and against the file drawer

Physiognomy structurally cannot perform the third discipline because it functions as a marketing posture rather than a falsifiable model. A model that predicts a graded ordering can fail. The ordering can come out flat, or inverted, or the predicted-strong axis can collapse while the predicted-weak axis survives. The discipline is to specify, in advance, what failure looks like. Then, if it happens, you must report it as a refutation rather than reach for the selected component, the larger subgroup, or the alternative estimator that would rescue the headline. The file drawer is how an entire subfield convinces itself of an effect that isn't there. Facial width-to-height ratio and aggression is the cautionary tale. The magnitude that lab studies pooled to was $d \approx 0.43$, the meta-analysis's own laboratory-only subset (Haselhuhn, Ormiston & Wong, 2015), versus an overall pooled $d \approx 0.22$. This collapsed when Kosinski (2017) tested fWHR against 55 constructs across 137,163 individuals and found no relationship. Reporting your own null carries credibility precisely because it is the act physiognomy never performs. It also guards against the use case. A result published as a refutation cannot be repurposed as a screening tool.

These three virtues are not independent. They interlock: pre-registration binds a reported null, turning it from a convenience into a commitment, and demographic control then renders that null interpretable by revealing that the surviving signal was actually a confound. The willingness to report the null is what keeps pre-registration and partialling from collapsing into theater. Implement

all three and the distinction between your program and physiognomy shifts from a claim about your intentions to a property of your record.

4. The worked example: a pre-registered facial hierarchy that failed

We did not reach this conclusion through argument alone. We got here by running the test on ourselves and losing. The companion empirical paper (Abdo, 2026b) reports the full pre-registered test, including its adversarial design, the SHA-256 hash chain, the rare-class power ceiling, and the demographic-shortcut analysis of the surviving signal. We summarize here only what the worked example needs.

The common-cause model above yields a specific, risky prediction (developed formally in the companion theory paper, Abdo, 2026a): on facial morphology, readability should order by hormonal organization. We operationalized three tiers from the prior endocrine literature. These include a sexually-dimorphic "sex-modality" axis (high organization), a "function" axis predicted to carry the largest sex-independent developmental readout, and low-organization "letter" axes predicted to read near chance. We predicted, directionally, that the function axis would lead, the sex-modality axis would be intermediate and largely attributable to sex, and the letter axes would sit near null.

A necessary disclosure precedes these figures, one that supports rather than undermines the discipline this Perspective advocates. The trait axes are not validated psychometric instruments. They are unvalidated community-assigned typings from the Objective Personality System, specifically "coins," which lack published criterion validity. Human typers could view the person's face during assignment through a behavioral channel, creating a possible face-to-label leakage path. This leakage inflates any face readability estimate. Consequently, the headline result is null-conservative. A signal that fails to appear despite a channel that should manufacture one is a strong null. However, this caveat complicates the interpretation of the single signal that survived. When we attribute the sex-modality result to a sex-correlated demographic shortcut below, we cannot fully exclude that the label itself was partly face-assigned. This mirrors the Wang and Kosinski critique this Perspective levels at others. The point of this disclosure is that a program hiding a face-to-label path while claiming to police a face-to-feature one has not drawn the bright line. It has crossed it.

We then proceeded with the three disciplines in sequence.

We pre-registered. Before touching the held-out identities, we committed the hypothesis, the estimator (a covariance-whitened linear discriminant on the full 40-dimensional facial shape vector - whole-shape, no per-component selection, eliminating the pick the principal component that worked degree of freedom), the covariates, and an explicit decision rule that named, in advance, what would count as a refutation: a flat or inverted ordering, or the predicted-strong axis failing to separate above chance. We froze the estimator to a byte-identical script (SHA-256 c22dcb07…) and registered the plan on a public timestamped registry, twice - once for an initial held-out cohort (OSF DOI 10.17605/OSF.IO/SMRXT, osf.io/smrxt, n = 83) and again for a larger one (OSF DOI 10.17605/OSF.IO/KJ2RW, osf.io/kj2rw, n = 549). The timing is the load-bearing part. On the same calendar date (2026-06-15), the OSF registration timestamp precedes the confirmatory-run timestamp, and the held-out cohorts were sequestered, fixed and disjoint by construction through exclusion from the 707-identity discovery set, with their features extracted hypothesis-blind and SHA-256-fingerprinted before any held-out effect size was computed. The OSF registration and the commit/run timestamps are the auditable record.

We controlled for demographics. Biological sex was read from a database record, never from the

image. The primary estimate was sex-residualized. We replicated within each sex; within-ethnicity replication held in the discovery analysis (within-White $d = -0.38$, $n = 166$), but the ethnicity covariate was never wired into the frozen confirmatory estimator, which we flag as a pre-registration deviation (the sex-residualized primary is unaffected).

We ran exactly that, and reported what came back. On the larger held-out cohort, which comprised 549 identities disjoint from the discovery set and nearly ten thousand photos, we ran the frozen estimator with no discovery-number rescue on CPU-only hardware. The predicted hierarchy did not hold. The function axis, predicted to lead, came out below a low-organization letter axis. It ranked third of the four axes, with a sex-residualized effect of d $\approx$ 0.21, a bootstrap confidence interval spanning zero, and a failure to survive multiple-comparison correction (p $\approx$ 0.17). After adjusting for head pose as well, it shrank further (d $\approx$ 0.15) and still did not separate above chance. A low-organization "letter" axis, predicted to sit near null, actually outperformed it and survived correction. The pre-registered confirmation condition, which required the function axis and the sex-modality axis to survive correction together, failed because the function axis did not survive.

A note on power is in order because the phrase "did not separate above chance" should not be interpreted as a definitive null result that rules out a small effect. The cohort was well-powered to test the discovery-magnitude claim. Specifically, the discovery's central estimate did not replicate, which renders its magnitude implausible. However, the same group split leaves the design underpowered for detecting a small effect. Consequently, a weak function effect of approximately 0.2 is neither confirmed nor excluded.[1] Crucially, the hierarchy refutation does not rely on the power of the function coin. The predicted ordering inverted, shifting to sex-modality greater than S_N letter greater than function. An inverted ordering refutes the model regardless of whether the function coin's own small effect is real.

According to our pre-registered decision rule, this constitutes a refutation. We therefore report it as such.

This is the part that earns the bright line. Two axes survived multiple-comparison correction: the **sex-modality axis** (p $\approx 10^{-5}$) and, against the model's own prediction, a low-organization **letter axis** (p $\approx$ 0.003). But only the sex-modality axis, the strongest coin and the axis most saturated by biological sex, had a bootstrap confidence interval that excluded zero. Both the function axis the model needed to lead and the surviving letter axis had CIs spanning zero. The *only* CI-robust facial signal in the entire test was a sex-correlated one. It survives only linear sex-residualization, so residual, such as nonlinear, sex-correlated structure is the most parsimonious read. This signal asserted itself exactly where the model predicted sex would dominate but failed to leave room for the sex-*independent* developmental signal the model needed to claim a trace. That a predicted-near-null letter axis nonetheless out-survived the predicted-strong function axis sharpens the refutation rather than blunting it. The hierarchy did not merely fail to confirm; it came out, in part, inverted. An earlier, tiny, underpowered run had thrown a large but *inverted* function magnitude, a below-chance held-out AUC of 0.217. The held-out discriminant separated the classes in the direction opposite to discovery. The powered run dissolved this result. That small-n run was itself logged as a refutation, not a confirmation. That is the file drawer in miniature, caught and reported

[1] Even though the overall sample size was $n = 549$, only 59 lead-Te identities were present at the function coin's group split, which capped the study's power. The design offered approximately 83% power to detect the discovery's effect size of $d \approx 0.45$, but it failed to do so. This result places a 95% ceiling at $d \approx 0.60$. While 0.45 lies just inside that ceiling, a well-powered test would more often than not have flagged it. The same split yielded only about 29% power for the observed effect. It also provided between 26% and 50% power for a true effect size of $d = 0.2$-0.3. A companion sensitivity analysis developed the minimum detectable effect of $d \approx 0.43$ at 80% power.

rather than buried. The small-n result that *appears* to flatter the hypothesis is exactly the one a disciplined program must distrust.

This is the argument made concrete. Physiognomy is the practice of seeing a face-based signal, declaring it a trait, and never asking whether it is just the demographic confound. We asked. We pre-committed the question, we removed sex from the image and from the representation, and when we did, our headline trait-readout collapsed and the only signal whose confidence interval excluded zero was sex itself, the precise failure mode physiognomy is built not to notice. We could have reported the underpowered run, or the selected component, or the within-subgroup slice that looked better, and called it a confirmation. The discipline is that we did not, because we said in advance that we would not.

A confirmed hierarchy would have been a strong result. A refuted one, reported honestly under pre-registration, is arguably the more valuable contribution to this specific literature. The scarce and credibility-bearing element in face-based research is not another positive claim. It is a demonstration that a researcher in this field will let the confound win when the confound is what is there. The null is the evidence that the bright line is real. A program that draws it will, when tested, sometimes find itself on the physiognomy side of its own prediction and say so.

5. Ethics: a prior is not a verdict

The discipline of the method must be matched by discipline about use, and here the conclusion is categorical rather than probabilistic.

A prior is not a verdict. Even the most robustly confirmed population-level ordering would license only a probabilistic prior - a starting distribution, the first and weakest word about a person, to be overwritten the moment that person does anything. Two facts compound to make that prior almost worthless at the level of an individual. The effect sizes in this entire neighborhood are modest ($|d|$ on the order of a few tenths even where real). And the proxy literatures that purport to read prenatal hormone exposure are themselves contested; the popular 2D:4D digit-ratio proxy, for instance, shows near-null associations with measured testosterone in well-powered meta-analysis (Sorokowski, Kowal, et al., 2024). So the prior is nearly uninformative and the person's actual behavior dominates almost immediately. A model that forbids individual prediction in principle, and whose effect sizes forbid it in practice, is not a screening instrument and cannot be honestly sold as one.

No individual prediction. No screening, ranking, or gating. The legitimate object of this work is *mechanism* - whether the face is, at the population level, a correlated trace of a developmental signal. It is not, and must not become, a method for deciding about a person: not for hiring, lending, admissions, policing, dating, insurance, or any allocation of opportunity or suspicion. This is not a disclaimer appended for safety; it follows from the structure of the claim. A common-cause trace says nothing about worth, ability, intent, or outcome, and any apparatus that converted it into a per-person judgment would have abandoned the model and rejoined physiognomy. The demographic confounds make this concrete: because the confounds track demographic groups, any individual-level use would concentrate its errors - and its harms - on exactly those groups, reproducing the oldest injury of the physiognomic tradition under a new coat of statistics.

The asymmetry that keeps it honest. Reporting one's own nulls is not only good epistemics; it is an ethical firewall. A result published as a refutation cannot be quietly relicensed as a product. The same discipline that protects the science from self-deception protects the public from the use

case. The discomfort a reader may feel at any face-and-disposition research is the correct response, and the appropriate answer to it is not reassurance but constraint: pre-register, control for the confounds, report the nulls, and refuse structurally in advance to make the individual prediction. Where a program will not accept those constraints, the discomfort is warranted and the work belongs with the pseudoscience.

The bright line is not a boundary between two topics but between two practices. One approach sees a face, asserts a character, sells a verdict, and never looks at the confound. The other proposes a shared developmental cause, predicts a graded ordering, pre-commits the test, removes the confound, and when the confound turns out to be all that was there, says so out loud. We ran the second practice on our own most-wanted result and it failed. Reporting that failure is the evidence we can offer that the line exists.

Data and Code Availability

The pre-registrations at osf.io/smrxt and osf.io/kj2rw contain timestamped SHA-256 fingerprints of the confirmatory artifacts, recorded before any held-out effect size existed: the held-out cohort CSV (e5c5786f…), the frozen feature arrays (2002e174…), and the frozen estimator (confirmatory_heldout.py, c22dcb07…). Upon publication, we will deposit the derived 3DMM shape features, the held-out cohort CSV, and the per-coin result tables (effect sizes, bootstrap CIs, and p-values, as detailed in Section 4) alongside the frozen estimator (confirmatory_heldout.py), as a release component within the same public OSF project that holds the two pre-registrations (osf.io/smrxt and osf.io/kj2rw). Source celebrity photographs are not redistributed, for rights reasons. Depositing the derived 3DMM shape features instead lets the reported null be re-derived from the released features and the fingerprinted estimator.

Competing Interests and Funding

The author declares no competing interests and received no external funding for this work.

Notes

References

Abdo, M. (2026a). Prenatal Hormonal Milieu as a Common Cause of Correlated Facial and Psychological Phenotype: A Multi-Axis Model and Its First Falsification Test. *PsyArXiv.* https://doi.org/10.31234/osf.io/akwrg

Abdo, M. (2026b). Testing a Prenatal-Hormone Effect-Size Hierarchy in Facial Morphology: A Pre-Registered Self-Refutation at Celebrity Scale. *PsyArXiv.* https://doi.org/10.31234/osf.io/mq4db

Agüera y Arcas, B., Mitchell, M., & Todorov, A. (2018). Do algorithms reveal sexual orientation or just expose our stereotypes? *Medium.* https://medium.com/@blaisea/do-algorithms-reveal-sexual-orientation-or-just-expose-our-stereotypes-d998fafdf477

Berenbaum, S. A., & Beltz, A. M. (2011). Sexual differentiation of human behavior: Effects of prenatal and pubertal organizational hormones. *Frontiers in Neuroendocrinology, 32*(2), 183-200. https://doi.org/10.1016/j.yfrne.2011.03.001

Haselhuhn, M. P., Ormiston, M. E., & Wong, E. M. (2015). Men's facial width-to-height ratio predicts aggression: A meta-analysis. *PLoS ONE, 10*(4), e0122637. https://doi.org/10.1371/journal.pone.0122637

Kosinski, M. (2017). Facial width-to-height ratio does not predict self-reported behavioral tendencies. *Psychological Science, 28*(11), 1675-1682. https://doi.org/10.1177/0956797617716929

Phoenix, C. H., Goy, R. W., Gerall, A. A., & Young, W. C. (1959). Organizing action of prenatally administered testosterone propionate on the tissues mediating mating behavior in the female guinea pig. *Endocrinology, 65*(3), 369-382. https://doi.org/10.1210/endo-65-3-369

Sorokowski, P., Kowal, M., et al. (2024). Relationship between the 2D:4D and prenatal testosterone, adult level testosterone, and testosterone change: Meta-analysis of 54 studies. *American Journal of Biological Anthropology, 183*(1), 20-38. https://doi.org/10.1002/ajpa.24852

Wang, Y., & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology, 114*(2), 246-257. https://doi.org/10.1037/pspa0000098

Whitehouse, A. J. O., Gilani, S. Z., Shafait, F., Mian, A., Tan, D. W., Maybery, M. T., Keelan, J. A., Hart, R., Handelsman, D. J., Goonawardene, M., & Eastwood, P. (2015). Prenatal testosterone exposure is related to sexually dimorphic facial morphology in adulthood. *Proceedings of the Royal Society B: Biological Sciences, 282*(1816), 20151351. https://doi.org/10.1098/rspb.2015.1351