Prenatal Hormonal Milieu as a Common Cause of Correlated Facial and Psychological Phenotype: A Multi-Axis Model and Its First Pre-Registered Falsification Test

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

Theoretical model with a pre-registered falsification test; a refutation, not a confirmation.

Abstract

Prenatal androgen exposure acts as an accepted common cause of co-occurring physical and behavioral traits - digit ratio, spatial cognition, and the anatomical-behavioral profile of congenital adrenal hyperplasia all trace to a shared hormonal gradient. We generalize this single-axis precedent into a **multi-axis common-cause model**: across critical fetal windows, the intrauterine hormonal milieu (testosterone, cortisol, thyroid, estradiol, and others, varying in dose and timing) simultaneously shapes physical structure, including the face, and neural architecture, so that observable form constitutes a *correlated trace* of the developmental event that shaped disposition rather than a cause of it. The model yields a falsifiable prediction absent from prior physiognomic claims: the **readability of a trait-axis from the face increases monotonically with the strength of that axis's prenatal hormonal organization.** We test this prediction against our own pre-registered, held-out data and report a refutation. A prior, non-pre-registered discovery analysis had produced an encouraging function-axis signal ($d \approx -0.47$, sex-residualized -0.45), so the held-out result reads most precisely as a **failed replication.** We pre-registered the prediction in two registrations (OSF DOI 10.17605/OSF.IO/SMRXT, $n = 83$; OSF DOI 10.17605/OSF.IO/KJ2RW, $n = 549$) and ran a frozen estimator on held-out cohorts the discovery analysis never touched. The predicted ordering did not hold: the well-powered test produced sex-modality $|d| = 0.39 >$ letter $|d| = 0.27 >$ function $|d| = 0.21$, with the flagship function axis failing to separate above chance and failing FDR correction while a letter axis predicted to lie near the null survived. The readability-ordering form of the model is therefore **not supported.** The value here is as methodological as theoretical: a physiognomy-adjacent claim made precise enough to be killed by its own pre-registered test.

---

1. Introduction

The idea that the face reveals character boasts a long history, though it has been thoroughly discredited. Classical physiognomy posited a direct body-to-mind causal arrow. This approach collapsed reliably on confounds such as image quality, expression, grooming, demographic correlates, and the self-fulfilling effects of being treated as one's appearance suggests. Modern machine-learning revivals of the genre inherit the same fatal structure. They predict a socially loaded label from a photograph. These systems cannot separate a genuine developmental signal from the artifacts that co-vary with it.

This paper asks whether a different causal structure can both ground the recurring observation that morphology and disposition co-vary and generate a prediction sharp enough to be falsified. We are talking about a model that explicitly denies the body to mind arrow. We propose such a

structure. We derive a falsifiable ordering prediction from it. We then report the result of testing that prediction against our own pre-registered held-out data.

Let us be clear about the paper's trajectory. We offer two things. First, a theoretical contribution: the model and the falsifiable hierarchy it generates. Second, an empirical contribution: a refutation of that hierarchy's central claim. Why present both? Because a falsifiable model that fails its first pre-registered test is a more honest scientific object than one decorated with post-hoc confirmation. What follows is simply a falsifiable model and its first failed pre-registered empirical test. This is a held-out replication attempt that did not reproduce an encouraging discovery-cohort signal (§4.1).

---

2. The Model

2.1 From single-axis precedent to a multi-axis generalization

The organizational hypothesis, which posits that perinatal hormones permanently and jointly shape both soma and brain, finds strong single-axis support in Berenbaum and Beltz (2011). In individuals with excess prenatal androgen due to CAH, virilization of anatomy occurs alongside a shift in behavior, as documented by Berenbaum and Resnick (1997). Similarly, opposite-sex twins display masculinized digit ratios and elevated sensation seeking, according to Cohen-Bendahan et al. (2005). Cord and amniotic testosterone have been shown to predict later language and autistic-trait measures, with Whitehouse et al. (2012) and Auyeung et al. (2009) providing this evidence. The common form across these results is one hormonal gradient yielding two correlated outputs: body and behavior.

We suggest that this structure extends beyond androgens and clinical extremes to encompass the typical case. Multiple hormone axes vary continuously in both concentration and the timing of exposure across critical windows, such as the masculinization programming window (Welsh et al., 2008). These axes jointly specify a high-dimensional developmental trajectory whose outputs include both morphology and disposition. We flag immediately that this generalization runs ahead of the evidence and will return to this point in §5. The strongest single-axis result, 2D:4D, is itself meta-analytically near-null even as a marker of testosterone (Sorokowski & Kowal, 2024; §4.4). Building a multi-axis edifice on this foundation demands caution.

2.2 Two coupled maps from one input

Imagine a fetus's developmental input as a vector h of hormone exposures, indexed by axis and critical window. The model proposes two coupled maps derived from that same input:

- a morphological map $\mathbf{M(h)} \rightarrow$ adult observable form (facial geometry, vocal-tract structure), and
- a neural map $\mathbf{N(h)} \rightarrow$ dispositional architecture.

Because **M** and **N** share the argument **h**, their outputs correlate without any causal arrow between body and mind. Personality is not read from the face; both are reads of **h**. This is the decisive departure from physiognomy, which posits body to mind causation and therefore collapses on confounds.

Two corollaries follow directly and match commonly reported observations:

- **Unrelated strangers can share a type** (same **h**-region → same M, N) across distinct genomes - convergent developmental expression rather than shared ancestry.
- **Identical twins can differ** (same genome, different intrauterine **h** via placental/cord/diffusion asymmetry → different M, N) - locating their difference in the non-shared developmental residual, the variance component that the companion twin-synthesis paper (Abdo, 2026a) identifies as the natural home for a prenatal developmental contribution and as the empirical motivation for this model.

2.3 The falsifiable prediction

A model earns the right to be taken seriously only if it forbids something. This one does. Define, for each trait-axis a, its hormonal organization O(a) - how strongly axis a is specified by h (strong for sexually dimorphic axes wired by large androgen gradients; weaker for axes with diffuse or late hormonal input). The model predicts:

Readability of axis a from facial morphology increases monotonically in $\mathbf{O}(a)$.

This is a hierarchy prediction, not a point prediction, and it is falsifiable three ways: (1) if readability is flat across axes, the common-cause coupling is absent; (2) if the ordering inverts, the model is wrong; (3) if a weakly organized axis reads strongly, an alternative cause (cultural, self-presentation, demographic artifact) is implicated. No prior physiognomic claim made an ordering prediction. They claimed uniform legibility of character, precisely the form that confounds counterfeit. An ordering that tracks independently known hormonal organization resists faking by image-quality or demographic artifacts, because those confounds do not respect O(a).

Operationally, we tiered three families of trait-axis by O(a) from the prior hormonal literature. We note that the high rank assigned to the function axis was not literature-only: a prior discovery analysis on this face dataset had returned an encouraging function-axis effect (§4.1), so the held-out test is a replication attempt, not a first look. The three families:

- **sex-modality** axis - sexually dimorphic, highest O(a);
- **function** axis - the model's flagship intermediate readout, predicted high;
- **letter** axes - predicted lowest O(a), at or near the null.

The predicted sex-residualized readability ordering was frozen before the test.

> **function $\geq$ sex-modality $>$ letter axes**, with function and sex-modality each separable above chance and the letter axes at or near the null.

We deliberately included the sex-modality axis (an axis essentially saturated by biological sex) as an internal check: the central worry for any face-based claim is that the apparatus merely re-reads sex. The model's distinctive bet was that the *function* axis would carry the largest readout *after* sex is residualized out, i.e., a real, sex-independent developmental signal sitting above the letter axes.

2.4 Scope and self-imposed limits

This model makes a claim about populations and mechanisms, not a license for individual judgment. It predicts a probabilistic prior rather than a verdict. It says nothing about worth, intelligence, criminality, or outcome. The organizational signal is anatomical. It is not a statistical proxy for demographics. This is why demographic controls are mandatory in any test. Any attempt to rank or gate individuals would misuse the model, not apply it. We return to this in §6.

3. The Falsification Test

We publicly pre-registered the prediction from §2.3 and tested it using a frozen estimator on held-out cohorts that the discovery analysis never touched. We report the result verbatim; it is a refutation. The full methodological treatment of this test-the adversarial design, the SHA-256 hash chain, the rare-class power ceiling, and the demographic-shortcut analysis of the surviving signal-is the subject of the companion empirical paper (Abdo, 2026b). We summarize here only what is needed to adjudicate the model's prediction.

3.1 Design

We registered the prediction on OSF on two occasions, both on 2026-06-15, prior to the confirmatory runs.

- **Registration 1** - DOI 10.17605/OSF.IO/SMRXT (osf.io/smrxt), held-out cohort n = 83 identities (1,364 photos, all fitted, per the smrxt confirmatory result document `confirmatory_heldout_results.md` - `heldout_pose.npz`, 1364/1364 fitted).
- **Registration 2 (powered replication)** - DOI 10.17605/OSF.IO/KJ2RW (osf.io/kj2rw), held-out cohort $n = 549$ identities (9,966 photos), all held-out identities with $\geq$3 model fits and a full 5-coin OPS stack, disjoint from the 707-identity discovery cohort. The estimator, hypothesis, and decision rule were unchanged from Registration 1; only the cohort expanded. (The *sex-residualized* primary - the §3.2 headline analysis - is computed on $n = 547$: 2 of the 549 identities, Ashley Banfield and Cody Talks, are dropped for missing sex in the annotation database, so the per-coin sex-residualized n's in §3.2 sum to 547, e.g. sensory_mf 304/243. The 549 cohort count and the 547 sex-residualized analysis count are thus not silently inconsistent.)

Where the larger cohort came from. The two registrations sequester two *nested* held-out pools, both defined purely by exclusion from the 707-identity discovery manifest. Registration 1 (smrxt) locked a first batch: the 83 identities with a complete 5-coin OPS stack among the 115 OPS-typed folders then identified as absent from the discovery manifest. The smrxt frozen-cohort record `osf_prereg_paper3.md` registers the count: 115 OPS-typed folders absent from discovery $\rightarrow$ 83 with a full 5-coin stack, the other 32 dropped for incomplete types. Registration 2 (kj2rw) widened the same exclusion rule to the *complete* set of OPS-typed identities absent from the 707-identity discovery cohort (every such identity with a full 5-coin type and $\geq$3 fitted photos), which numbers 549. The additional ~466 identities are not a new construction but the remainder of that same disjoint-by-construction pool, assembled after the smrxt registration. We do not claim the full 549 list was locked at the smrxt stage. It was not. We tie its sequestration to the kj2rw timestamp instead. The powered registration (osf.io/kj2rw §4, §6) records that all 549 identities were frozen feature-blind at registration. Hypothesis-blind 3DMM and pose features extracted for the cohort had **no** per-coin effect size, Cohen's d, LDA, or hypothesis test computed at registration time. This binds the frozen cohort list, blind features, and estimator under SHA-256 fingerprints (`heldout_large_cohort_frozen.csv` `e5c5786…`; features `2002e174…`; estimator `c22dcb07…`) recorded at the registration timestamp, so any later claim that the cohort was assembled after seeing held-out effects is falsifiable by re-hashing.

Sequestration order (auditable). The held-out identities were sequestered, fixed and disjoint by construction, before any held-out feature was loaded because both cohorts were defined by

exclusion from the 707-identity discovery cohort. The estimator was committed as a frozen script, which froze the hypothesis and decision rule. Two distinct fingerprints anchor the two registrations. The smrxt registration sealed the estimator helper `confirmatory_analysis.py` (SHA-256 `578ec8f4…`; osf_prereg_paper3.md §6-§7). The kj2rw registration sealed the held-out runner `confirmatory_heldout.py` (SHA-256 `c22dcb07…`, §"Data and Code Availability"; osf_prereg_paper3_powered.md §3, §6), which applies that helper unchanged. The runner (`c22dcb07…`) was applied byte-identically across both cohorts. This means "identical across both runs" refers to the runner being applied unchanged to each cohort, not to a single hash registered at both timestamps. On 2026-06-15 the OSF registration timestamp preceded the confirmatory run. The powered result file records this as a "post-registration timestamp." We note the registration and run share that single date and present this order as the basis for the "before the confirmatory runs" claim. The OSF and commit timestamps, along with the §6 artifact hashes recorded at registration, are the auditable record.

The estimator remained frozen and identical across both runs. It used the full 40-dimensional 3D-morphable-model (BFM, 3DDFA_V2) per-identity shape centroid, with pose and expression dimensions discarded and **no** principal-component selection. We calculated covariance-whitened LDA / Mahalanobis separability per axis as a Cohen's-*d*-equivalent. Identity-disjoint GroupKFold (k = 5) ensured folds were leakage-free by construction. **Sex-residualization** served as the primary adjustment, projecting onto the null space of the linear biological-sex direction, with sex read from the annotation database - never from the image. Head pose (yaw/pitch/roll) was regressed out as a confound. We computed bootstrap 95% CIs using 1,000 identity resamples, along with Benjamini-Hochberg FDR (q = 0.05) across axes. The decision rule was frozen in advance. The §2.3 ordering holding *with both function and sex-modality surviving FDR* confirms. A flat/inverted ordering, or a function axis that fails to separate above chance after sex + pose adjustment, refutes. The registrations committed in advance to reporting the held-out result verbatim, with no substitution of discovery (selected-PC) numbers and no estimator re-selection.

3.2 Result

The well-powered sex-residualized readability ordering involved a sample of n = 549.

Rank	Axis		d
1	sex-modality	**0.39** (AUC 0.609)	**survives** ($p = 1.1 \times 10^{-5}$)
2	letter (S_N)	0.27 (AUC 0.574)	**survives** (p = 0.0026)
3	function	0.21 (AUC 0.560)	fails (p = 0.172)
4	letter (F_T)	0.12 (AUC 0.534)	fails (p = 0.169)

- **Predicted ordering:** function ≥ sex-modality > letters.
- **Observed ordering:** sex-modality > letter (S_N) > function > letter (F_T).

The flagship function axis ranked below both the sex-modality axis and a letter axis. Its whole-shape bootstrap CI [−0.377, +0.604] spans zero around the resampled point estimate d = +0.13; note that the d = +0.21 above is the GroupKFold whitened-LDA estimate on the same axis. Because the interval spans zero the axis is not separable above chance. After sex + pose adjustment its effect is only d = +0.15. The sex-residualized AUC is 0.560 and no pose-adjusted AUC was computed for this axis. It fails FDR with p = 0.172. The pre-registered refutation condition is met. This

condition is the §3.1 decision rule: a function coin that fails to separate above chance after sex + pose adjustment and corresponds to OSF §5. Meanwhile a letter axis (S_N, predicted near-null) survives FDR. This result is the opposite of the hierarchy.

The sex-modality axis behaved exactly as the model's high-rank prediction expected. The strongest coin yielded a d of +0.39 using GroupKFold whitened-LDA with an AUC of 0.609, an FDR p of 1.1×10^{-5}, and a bootstrap CI of [+0.003, +0.604] that excluded zero. This CI is around the resampled point estimate d of +0.31. We report this for completeness, but it does not rescue the verdict. This axis is precisely the readout the model needed to exceed, not merely match. A candidate explanation, which we offer as a hypothesis rather than an established interpretation, is that the surviving signal is plausibly dominated by residual sex covariance that survives linear sex-residualization. Sex-residualization projects out only the linear sex direction. We do not treat this as demonstrated. It sits in tension with the independent position that the sex-modality axis encodes cognitive modality rather than biological sex.

The smaller registration ($n = 83$) also refuted the hypothesis, although it was underpowered on the function axis (10 Se / 12 Te). There, the function effect was inverted (sex-residualized AUC 0.217, below chance), sex-modality collapsed to last ($|d| = 0.145$), and no coin survived FDR. The two registrations agree on the verdict for different reasons, namely a small-n artifact in the first and a well-powered null in the second, and the function axis fails the confirmation condition in both.

3.3 The verdict, stated plainly

Both registrations return REFUTES according to the frozen pre-registered decision rule. The model's central falsifiable prediction failed in this first pre-registered test (§4.1). That prediction was that the function axis carries the largest sex-independent prenatal-hormone readout, sitting above the letter axes. Instead, a letter axis the model placed at the null survived. On the evidence we have, the readability-ordering form of the model is not supported.

---

4. Discussion

4.1 The discovery signal this test failed to replicate

To ensure full transparency regarding the empirical scaffolding behind the function prediction, we note that before the pre-registered held-out test, a non-pre-registered discovery analysis on the same face dataset had produced an encouraging function-axis result. On the 707-identity discovery cohort, a function-axis coin (observer/decider) showed d = −0.47 that survived biological-sex adjustment (sex-residualized d = −0.45). This effect replicated within each sex (males −0.57, females −0.37) and within ethnicity (within-White d = −0.38 at full ethnicity coverage, n = 166). The larger −0.62 came from the low-coverage n = 40 subset before full FairFace labeling, and the source flags the full-coverage figure as resolving that coverage caveat. It was characterized as a "large" discovery effect per the selected-PC discovery comparison in the powered-results file. That discovery signal, not literature alone, is why the model assigned the function axis its high rank. The pre-registered held-out result reported in §3 is therefore most honestly read as a failed replication. On the well-powered held-out cohort the same function axis collapsed to a near-zero, non-significant effect (sex-residualized d = +0.21, pose-adjusted +0.15, bootstrap CI spanning zero, FDR p = 0.172). We disclose the discovery confirmation here precisely so the refutation is not overstated as a first-ever look and the discovery scaffolding is not concealed. The failure of an encouraging discovery signal

to replicate on sequestered, pre-registered data is the strongest and most honest form of the result.

4.2 What the refutation kills

The refutation is specific and real: the monotone-in-O(a) readability hierarchy is not supported. After the pre-registered adjustments, the function axis, which is the part of the model that distinguished it from "the face reads sex," does not separate above chance, while a letter axis that the model demoted to the null does. An honest reading is that the specific ordering the model generated is wrong, at least as operationalized here. We do not get to keep the hierarchy.

This matters because the hierarchy was the model's entire claim to non-triviality. Without it, what remains?

4.3 What, if anything, survives

Two things are worth separating.

The common-cause logic versus the specific ordering. The model consists of two layers: (a) a *structural* claim that body and disposition correlate because both read a shared developmental input **h** (no body→mind arrow), and (b) a *quantitative* claim that face-readability orders monotonically by hormonal organization. We refute layer (b) here. Layer (a) is *not directly tested* by this experiment - it is the prior that motivated the ordering, and the failure of (b) neither confirms nor refutes it. What the failure does establish is that, if (a) is true, the mapping from "hormonal organization" to "facial readability" is not the simple monotone function the model assumed: the function axis's hormonal organization (as we tiered it from the literature) did not translate into facial separability. The model's weakest link may be the O(a) tiering itself - derived from a contested literature - rather than the common-cause structure.

The surviving signal, soberly. One axis (sex-modality) reads robustly from facial shape after sex-residualization, and one letter axis (S_N) reads weakly but FDR-significantly. That two axes separate above chance at all establishes only that *some* facial-shape signal exists for *some* trait-axes - a far weaker statement than the model made, and one fully consistent with mundane explanations (residual sex covariance for the sex-modality axis; demographic or self-presentation correlates for S_N). A further alternative explanation applies to *both* survivors: a **face→label leakage path**. The companion empirical paper (Abdo, 2026b) flags that OPS labels are assigned by human typers reasoning over behavioral *video* - i.e., the typers see the person's face while assigning the coin (that paper names this "a possible face→label leakage path"). Any such leakage would *inflate* a positive face-readability estimate, so it is a material alternative explanation for precisely the two positive coins that survive here, further weakening the case that the survivors represent independent confirmation. (It does not bear on the function-axis null: leakage can inflate a positive effect but cannot manufacture a null, so the flagship refutation stands.) We explicitly decline to spin the two surviving coins into a partial confirmation. The pre-registered confirmation required function *and* sex-modality surviving FDR; function did not.

4.4 The contested mechanism literature

Intellectual honesty demands that we highlight the contested nature of the model's mechanistic foundation. This reality weakens the prior, regardless of our test.

- **2D:4D digit ratio.** The single most-cited proxy for prenatal androgen exposure (Manning et al., 1998) is itself weakly validated at its first link: a meta-analysis of 54 studies finds 2D:4D

essentially unrelated to testosterone (Sorokowski & Kowal, 2024), undercutting its standing as a marker of prenatal androgen exposure **h**. Its onward associations with behavioral *outcomes* are likewise small: a meta-analysis of 2D:4D and aggression recovers only $r \approx -0.06$ in men (Hönekopp & Watson, 2011). If the flagship physical proxy for **h** is this weakly tied to androgen and to downstream behavior, a strong multi-axis body↔mind coupling is *a priori* less secure - a direction consistent with our function-axis null.

- **Extreme-male-brain / fetal-testosterone framing.** The broader organizational account as applied to neurotypical personality remains disputed: the largest test of the empathizing-systemizing and extreme-male-brain predictions recovers only modest sex differences and a partial autistic-trait pattern (Greenberg et al., 2018), and debate over effect sizes and causal interpretation of amniotic-testosterone→trait links continues.

The refutation comes as no surprise. It aligns with a mechanistic literature that is already weak and contested. We offer the model because its form constitutes a genuine contribution. This form features a falsifiable ordering and an explicit non-physiognomic causal structure. We do not present it because its mechanistic premises are secure.

4.5 Why publish a refuted model

Three reasons. First, the model's form is the contribution. It takes a vague, confound-prone, physiognomy-adjacent intuition and converts it into an ordering prediction. A held-out, pre-registered test could kill this prediction, and it did. Most claims in this space are never made falsifiable enough to fail. Second, the negative result is informative. It tells the field that the obvious developmental signal beyond sex does not fall out of facial shape under clean adjustment. This should temper the next round of face-based personality claims. Third, transparency about a failed flagship prediction is the only defensible way to discuss a topic with physiognomy's history. Presenting the model with its refutation is the safeguard.

---

5. Limitations

- **The function-axis null may reflect the operationalization, not the model.** Whole-shape BFM centroids, the specific LDA/Mahalanobis separability estimator, and the in-the-wild photo source each impose ceilings; a different morphological representation could in principle recover a signal. We pre-registered *this* estimator and report *its* verdict, and we do not claim a different estimator would have confirmed.
- **The O(a) tiering is the model's softest joint.** It was derived from a contested hormonal literature before the test, but if that literature is wrong about which axes are "loud," the ordering prediction inherits the error. The refutation cannot distinguish "common-cause structure is wrong" from "O(a) tiering is wrong."
- **Ethnicity covariate not applied in the confirmatory path.** Cached FairFace labels covered 466/549 held-out identities, but the frozen estimator never wired ethnicity into the confirmatory computation (identical to the $n = 83$ run); no scraping was performed, as the pre-registration forbids it. The sex-residualized primary is unaffected, but full demographic adjustment remains incomplete. Reported as a deviation in both result documents.
- **Pose proxy deviation.** The frozen 5th pose covariate used within-identity pose dispersion (|yaw| std) rather than the discovery within-identity shape-dispersion proxy; yaw/pitch/roll (the actual pose adjustment) are identical to pre-registration. Preserved unchanged to keep

the estimator byte-identical, not corrected. Reported as a deviation.

- **Self-reported / proprietary typing as ground truth.** The trait-axis labels derive from an OPS-style typing system whose construct validity is itself unestablished; measurement error in the labels would attenuate any true signal and is not separable from a genuine null here.
- **One morphological modality, one data regime.** Celebrity-scale in-the-wild photos; no vocal, longitudinal, or biomarker data. The model spans more than the face; this test interrogates only the facial map.
- **The surviving coins are not independent confirmation.** The sex-modality survivor is plausibly residual sex; the S_N survivor is unexplained and could be a confound. Neither was predicted by the hierarchy, and the face→label leakage path (§4.3) could further inflate either positive coin.

---

6. Ethics

This work engages with the most dangerous claim in applied psychology, the idea that the face reveals the person. It does so to constrain, not license, that claim.

- **Common cause ≠ body→mind.** The model's whole point is to deny the physiognomic arrow. Body and disposition correlate, under the model, only because both read a shared prenatal input; neither causes the other. No facial measurement, under this model, *determines* a trait.
- **A prior, never a verdict.** Even where a signal survives, it is a weak, population-level probabilistic prior, not an individual judgment. The model is explicitly forbidden from ranking, gating, screening, or sorting individuals; any such use is a misuse.
- **The refutation reinforces the ethics.** The flagship sex-independent readout *failed.* The empirical takeaway is precisely the anti-physiognomic one: there is no clean "personality from face" signal beyond sex in these data. We publish the negative result specifically so it cannot be cited as support for face-based screening.
- **No deployment.** No classifier, score, or product follows from this paper. The artifacts are pre-registrations and a refutation.

Competing Interests

The author developed, and holds a continuing interest in, the proprietary OPS-style typing system from which the trait-axis labels used here are derived. That system is the locus of the author's broader applied work and represents a potential commercial stake. This interest is material to a physiognomy-adjacent claim, so we disclose it explicitly; we note that the result reported here is a **refutation** of the model's flagship prediction, i.e., a finding that runs against, not toward, any commercial interest in face-based typing. No other competing interests are declared.

Funding

This work received no external funding and was self-funded by the author. No funder had any role in the study design, analysis, interpretation, decision to publish, or preparation of the manuscript.

Research-Conduct and Data Provenance

This study is a secondary analysis of publicly available, third-party, in-the-wild celebrity face images. No images were generated, staged, or solicited for this study. No human-subjects intervention or interaction occurred. No identifiable private information beyond what is already public was collected. No consent was obtained because the analysis uses only pre-existing public images, and individuals are referenced by identity rather than redistributed. On this basis, secondary analysis of public, non-interventional data, and as an independent researcher with no affiliated institution and therefore no IRB of record, the author judges that no IRB review was applicable to this work rather than self-granting an institutional exemption. No IRB review was sought. The trait-axis (typing) labels are author-assigned via the proprietary system described under Competing Interests, not solicited from the depicted individuals.

Data and Code Availability

A frozen, pre-registered estimator was used for the confirmatory analysis. We identify it here to ensure the claim that we did not re-select the estimator is auditable:

- **Frozen estimator:** `faces/deca_exploratory/confirmatory_heldout.py`, SHA-256 `c22dcb076658bca3729c05554dc2c6dbe6dab0a56384111e86485fa42a548df3`, the held-out runner that was applied byte-identically to the $n = 83$ (osf.io/smrxt) and $n = 549$ (osf.io/kj2rw) cohorts. Two distinct fingerprints were registered across the two timestamps and should not be conflated: the smrxt registration sealed the estimator *helper* `faces/deca_exploratory/confirmatory_analysis.py` (SHA-256 `578ec8f4a3bff66a85d2fbac11472294fcd175a557854bdb7c554c9cb381308f`; osf_prereg_paper3.md §6-§7), and the kj2rw registration sealed the held-out *runner* `confirmatory_heldout.py` (`c22dcb07…`; osf_prereg_paper3_powered.md §3, §6), which applies that helper unchanged (the $n = 83$ result file records `confirmatory_analysis.py` as "applied UNCHANGED … via `confirmatory_heldout.py`"). So "byte-identical between runs" denotes the runner being applied identically across both cohorts, not a single hash registered at both timestamps. The $n = 549$ run applied this runner unchanged via a thin staging wrapper (`faces/deca_exploratory/run_powered_confirmatory.py`) that only renames the larger feature/pose arrays into the two input filenames the frozen script loads; no estimator helper was touched.
- **Held-out feature and pose arrays:** the per-identity whole-shape (40-dim BFM) centroid features and head-pose covariates for the held-out cohorts, plus the frozen sex/label CSV (`heldout_large_cohort_frozen.csv`, SHA-256 `e5c5786…`), are deposited alongside the analysis code at the locations linked from the OSF registrations.
- **Restrictions:** the OPS-style trait-axis labels are proprietary; raw labels are released only as the derived binary coin assignments needed to reproduce the reported separability, not as the underlying typing system. Source photographs are third-party (in-the-wild celebrity images) and are referenced by identity rather than redistributed.
- **Pre-registrations:** the pre-registered hypotheses and frozen decision rules are recorded at OSF DOI 10.17605/OSF.IO/SMRXT ($n = 83$) and OSF DOI 10.17605/OSF.IO/KJ2RW ($n = 549$).

The analysis code, the frozen estimator script and hash, and the held-out arrays form the locus of reproducibility, while the OSF registrations serve as the locus for the pre-registered hypotheses

and decision rules.

References

- Auyeung, B., Baron-Cohen, S., Ashwin, E., Knickmeyer, R., Taylor, K., & Hackett, G. (2009). Fetal testosterone and autistic traits. *British Journal of Psychology, 100*(1), 1-22. https://doi.org/10.1348/000712608X311731
- 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
- Berenbaum, S. A., & Resnick, S. M. (1997). Early androgen effects on aggression in children and adults with congenital adrenal hyperplasia. *Psychoneuroendocrinology, 22*(7), 505-515. https://doi.org/10.1016/S0306-4530(97)00049-8
- Cohen-Bendahan, C. C. C., van de Beek, C., & Berenbaum, S. A. (2005). Prenatal sex hormone effects on child and adult sex-typed behavior: Methods and findings. *Neuroscience & Biobehavioral Reviews, 29*(2), 353-384. https://doi.org/10.1016/j.neubiorev.2004.11.004
- Greenberg, D. M., Warrier, V., Allison, C., & Baron-Cohen, S. (2018). Testing the Empathizing-Systemizing theory of sex differences and the Extreme Male Brain theory of autism in half a million people. *Proceedings of the National Academy of Sciences, 115*(48), 12152-12157. https://doi.org/10.1073/pnas.1811032115
- Hönekopp, J., & Watson, S. (2011). Meta-analysis of the relationship between digit-ratio 2D:4D and aggression. *Personality and Individual Differences, 51*(4), 381-386. https://doi.org/10.1016/j.paid.2010.05.003
- Manning, J. T., Scutt, D., Wilson, J., & Lewis-Jones, D. I. (1998). The ratio of 2nd to 4th digit length: A predictor of sperm numbers and concentrations of testosterone, luteinizing hormone and oestrogen. *Human Reproduction, 13*(11), 3000-3004. https://doi.org/10.1093/humrep/13.11.3000
- Sorokowski, P., & Kowal, M. (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*(3), e24852. https://doi.org/10.1002/ajpa.24852
- Welsh, M., Saunders, P. T. K., Fisken, M., Scott, H. M., Hutchison, G. R., Smith, L. B., & Sharpe, R. M. (2008). Identification in rats of a programming window for reproductive tract masculinization, disruption of which leads to hypospadias and cryptorchidism. *Journal of Clinical Investigation, 118*(4), 1479-1490. https://doi.org/10.1172/JCI34241
- Whitehouse, A. J. O., Mattes, E., Maybery, M. T., Sawyer, M. G., Jacoby, P., Keelan, J. A., & Hickey, M. (2012). Sex-specific associations between umbilical cord blood testosterone levels and language delay in early childhood. *Journal of Child Psychology and Psychiatry, 53*(7), 726-734. https://doi.org/10.1111/j.1469-7610.2011.02523.x

Pre-registrations (this work): - OSF DOI 10.17605/OSF.IO/SMRXT - held-out confirmatory test ($n = 83$), registered 2026-06-15. - OSF DOI 10.17605/OSF.IO/KJ2RW - powered replication ($n = 549$), registered 2026-06-15.

Companion papers (this series): - Abdo, M. (2026a). The Non-Shared Developmental Residual: What Twin Variance Implies for a Prenatal-Hormone Account of Correlated Phenotype. PsyArXiv. https://doi.org/10.31234/osf.io/qnmtg - 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 - Abdo, M. (2026c). Reading the Receipt, Not the Person: An Anti-Physiognomy Framing for Developmental Trace Inference. PsyArXiv. https://doi.org/10.31234/osf.io/2rbqx

---