

**Will Before Representation:
Gender Identity as Preference Over Trajectories, and What
Behaviour Cannot Certify**

PIATRA . INSTITUTE

August 2026

Abstract

Clinical definitions of transgender identity rest on an undefined notion of experienced gender. Recent philosophy grounds gender identity in a conative state, either an intrinsic desire for category membership (Howard and Laskowski 2026) or a judgment of subjective fit (Cosker-Rowland 2024). We formalise that proposal in active inference, where gendered mismatch can sit either in a belief about the present body or in a distribution over preferred future observations. The belief placement makes mismatch and misperception the same quantity: one nat of mismatch requires a posterior displaced by 1.00 sensory standard deviations and four nats require 2.65; mismatch falls to 0.35 of baseline when sensory noise triples; and it survives twenty years of self-observation only below an update rate of 3.6 per million exposures. The preference placement produces any mismatch with an unbiased body estimate, no relief under blur and no extinction. Behaviour is a weak channel for the preference. As social threat rises, the best achievable classifier of latent orientation falls from an AUROC of 0.85 to 0.50 on one observed policy and from 0.996 to 0.55 on twelve, and 0.3 percent of the class information survives. Pairs that a scalar score cannot separate differ in the domain needing action 0.867 of the time, against a base rate of 0.868. Under threat, exploration shifts from social to private probes, the private-to-social ratio rises from 0.93 to 172, and the fraction of agents whose self-model resolves falls from 1.00 to 0.27. The model makes no claim about any person or community, and within it an authenticity test is uninformative.

Definitions of gender identity

ICD-11 defines gender incongruence as a marked and persistent incongruence between an individual's experienced gender and the assigned sex, and places it outside the mental-disorders chapter (Reed et al. 2016). The definition is administratively clear and explains little, because it rests on *experienced gender*, which is usually glossed as the gender one experiences oneself as being. The explanans contains the explanandum.

Metaphysical theories may contain primitives; pain and phenomenal colour resist reduction without that being fatal. Experienced gender, however, is asked to do explanatory work in clinical documents, in law and in argument, and a primitive cannot do that work. Asked why a person seeks a different voice, the answer that they experience their gender as female restates the question.

Two recent accounts give the concept more structure. Howard and Laskowski (2026) separate the question of what constitutes gender categories from the question of what makes a person a member of one, answer only the second, and answer it with desire: a person belongs to the gender category to which they intrinsically desire to belong. Cosker-Rowland (2024) grounds identity in judgments of subjective fit, so that having gender G as part of one's identity consists in not taking G to fail to fit oneself. She argues that this matches trans people's testimony that they have always had the identity they have, and that it explains why gender identities merit respect, since normative experiences and judgments can merit respect. Both accounts improve on the norm-relevancy account of Jenkins (2018), and both keep first-person report where Bettcher (2009) placed it, as the default and not as evidence to be weighed.

Hernandez and Bell (2025) argue that the enterprise is misconceived. They trace the concept of gender identity to trans-antagonistic medical theory and the wrong-body model, argue that its main pragmatic function is explaining trans people to non-trans people in order to secure access to rights and care, and urge theorists to stop trying to substantiate it. We do not answer that objection, and our second result confirms part of it: if behaviour is required to certify an underlying orientation, the requirement cannot be met, and it fails most completely where it is pressed hardest.

The model concerns the individual's conative organisation. It says nothing about what constitutes the category *woman*, which has its own literature (Haslanger 2000). Any account that grounds identity in an orientation toward *G* must specify *G* without presupposing the answer, and the model reduces that residual problem without removing it.

Schopenhauer's will as a precedent

Schopenhauer's will lies beneath deliberation. It is a blind, pre-reflective striving that reaches consciousness only as representation, and his claim most relevant here is that one's own body is given twice, externally as representation and internally as will (Schopenhauer [1818] 2010; Wicks 2021). Two elements carry over.

The first is an ordering. Conative organisation precedes the concept of it, so a person can have durable preferences about their body and social position before acquiring the words to state them. Learning a word does not create the orientation, and the account must allow the orientation to exist before the label. Cosker-Rowland (2024) builds her account to preserve this property.

The second is an interface. Preferences over embodiment are enacted through the sensorimotor and interoceptive loop that also presents the body as an object, which is where control-oriented interoceptive inference locates selfhood (Seth and Tsakiris 2018).

The metaphysics does not carry over. Nothing computed below bears on whether will is the noumenal essence of the world. The formal work needs a functional role, so the model uses *conative* for the object it manipulates and reserves *will* for the ordering claim. Influence also runs in a direction Schopenhauer does not allow: labels, communities and experimentation reorganise attention and preference over time. Hirsh (2026) formalises this as a bidirectional loop between semantic self-representation and episodic experience, and it makes the slow parameter of our model revisable.

Preference and belief in active inference

Discussions of gender and predictive processing often use *prior* for several distinct objects, and the argument depends on separating them. In discrete active inference the generative model contains a prior over hidden states, a likelihood mapping states to observations, and transition probabilities under each policy. Separately, the agent has a distribution over preferred outcomes, and policies are scored by expected free energy against it (Da Costa et al. 2020; Parr et al. 2022). Precision is a further object, and attention is the allocation of precision (Feldman and Friston 2010).

object	question it answers
state prior $P(s_t)$	what state is likely now
likelihood $P(o_t s_t)$	what would that state look like
preferred outcomes $P^*(o_{t:T})$	which futures are wanted
precision Π	how much should this signal weigh
policy prior $E(\pi)$	what is habitual

A trans woman at a mirror may hold $Q(\text{facial hair present} | o_t) \approx 1$, an accurate posterior, while P^* places almost all its mass on futures without facial hair. The two distributions are over different things and do not conflict. The relevant discrepancy lies between the trajectory expected under the status quo and the preferred trajectory, and in active inference it is the risk term of expected free energy,

$$G(\pi) \approx \underbrace{D_{\text{KL}}[Q(o_\tau | \pi) \| P_w^*(o_\tau)]}_{\text{risk}} - \underbrace{I_Q(s_\tau; o_\tau | \pi)}_{\text{epistemic value}},$$

where the second term rewards policies that resolve uncertainty even when they deliver no preferred outcome (Friston et al. 2015).

The one existing free-energy proposal about gender leaves this choice open. Langer (2022) theorises gender from the free energy principle and connects it to trans phantoms, and the embodiment intuition of that proposal is retained here. A formal version must decide which object carries the mismatch, and the next sections compute the cost of the belief option.

Two constraints come from the affect literature. Free energy is a functional used in approximate inference and is not itself suffering. In deep active inference, valence requires higher-order beliefs about how well the model is performing (Hesp et al. 2021), or tracks the rate of change of free energy (Joffily and Coricelli 2013). Dysphoria must therefore be derived from divergence together with importance, attention, controllability, expected duration and social threat, and cannot be identified with the divergence alone. Euphoria is likewise not the negative of that quantity; community accounts give it its own content, including affirmation and recognition (Beischel et al. 2022).

The resulting model has two timescales. On the slow timescale a conative organisation w structures P_w^* over embodied and social trajectories; an identity label ℓ is a culturally mediated representation of w through the categories and affordances N available in a person’s setting; and N is enacted as well as believed (Albarracin and Poirier 2022). On the fast timescale the hidden state comprises morphology, voice, endocrine and interoceptive condition, social recognition, relational position and enacted role; observations arrive in the corresponding channels; and policies include presentation, name and pronoun use, voice work, disclosure, medical intervention, inaction and reversible experiment. Preferences factorise across domains with individual weights,

$$P_w^*(\tau) \propto \prod_k P_k^*(o_\tau^{(k)})^{\alpha_k},$$

so that one person’s weight may fall on morphology and voice and another’s on social recognition and legal classification, with no stereotyped package required of either. Incongruence is the divergence between P_w^* and the trajectory conventionally predicted from the assigned category. It describes a latent structure and sets no threshold for deciding who counts.

Distress, persistence, euphoria and nonbinary orientations

The divergence between expected and preferred trajectories is one quantity. The negative affect attached to it is another, modulated by the weight of the domain, the attention it receives, its apparent controllability and its expected duration. A person can hold a durable divergence with little distress. Accounts that make distress constitutive must treat that person as an anomaly or as not trans; in the present model the two quantities are distinct from the outset.

The same separation covers the period after transition. Reducing the mismatch toward zero leaves P_w^* unchanged, because the update law contains no route by which a preference is revised when the world comes to match it. This is the structural basis of the persistence result below. A person with no remaining dysphoria retains the orientation that transition served, and a model whose object vanishes along with the distress has lost what it was meant to describe.

Euphoria has its own content. Community accounts describe self-affirmation, social affirmation and the discovery that something is available at all (Beischel et al. 2022). In the model these are three events: a preferred outcome realised, a social observation that moves the recognition domain, and a reduction in uncertainty about the reachable set, which is a payout of the epistemic term with no change in risk.

A factorised preference has no midpoint. Nonbinary and agender orientations appear as patterns of weight and acceptable range, and do not correspond to positions on a line between two endpoints (Matsuno and Budge 2017). The cost of ignoring this is computed below: an instrument that never asks about gender salience ranks a low-classification configuration at the 8th percentile of the mismatch it claims to measure. Genderfluidity corresponds to variation in the preferred region, or in its precision, across context and time, which the two-timescale structure permits.

Cisgender indifference corresponds to the width of the acceptable region. When the assigned trajectory lies comfortably inside it, no intense positive desire is required, so a desire-based account does not imply that cis people must want their gender strongly. People differ in the width and location of the acceptable region, and many never have occasion to find its edges.

Perceptual consequences of a belief architecture

Consider one body-feature dimension, scaled so that the standard deviation of its sensory observation is 1, so that displacements are in units a psychophysicist would report. A belief-first agent

holds a prior over the *current* value of the feature centred Δ away from the truth, with precision ratio r relative to sensory precision and $k = r/(1 + r)$. Its posterior estimate of its own present body is displaced by $b = k\Delta$, and the precision-weighted sensory prediction error remaining after inference is

$$D = \frac{1}{2} k^2 (\Delta^2 + 1) = \frac{1}{2} (b^2 + k^2).$$

Mismatch and misperception are therefore the same quantity. Since $k < 1$, any mismatch above half a nat forces $b \geq \sqrt{2D - 1}$, and numerical minimisation of the bias over the whole (k, Δ) grid reproduces that bound to within 0.001. One nat of mismatch requires a bias of 1.00 sensory standard deviations, two nats 1.73, four nats 2.65 and eight nats 3.87. A preference-first agent, whose divergence lies in P^* and never enters the state estimate, produces any of these levels with zero bias.

A bias of one standard deviation is a systematic error in judging one's own body large enough for ordinary psychophysics to detect, and the belief architecture requires it in every person with large mismatch.

The two architectures also separate under a manipulation any laboratory can perform: degrade the sensory channel and hold the belief fixed. Tripling the sensory noise reduces the belief agent's mismatch to 0.35 of baseline while its perceptual bias *rises* from 2.00 to 2.43 standard deviations, because a weaker observation is overruled more completely. Over the same range the preference agent's mismatch rises to 1.26 of baseline, and the mean component of its divergence, which reflects where the trajectory is heading independently of uncertainty, does not change. Blurring the mirror should relieve the belief agent and not the preference agent.

The largest difference appears over time. A prior about the present body is updated by observations of the present body, and a body seen twice a day for twenty years supplies 14,600 disconfirmations. At a per-exposure learning rate of one in a thousand the mismatch has a half-life of 0.47 years; at one in ten thousand, 4.75 years. Retaining nine tenths of the mismatch over twenty years requires an update rate below 3.6 per million exposures, about 28,000 times more rigid than the rate at which a body-representation prior is known to move within a single laboratory session: a full-body illusion shifted gendered self-identification within one sitting in 140 participants (Tacikowski et al. 2020). A belief that insensitive to evidence no longer functions as a belief.

The preference architecture predicts persistence for a structural reason, independent of any tuned parameter: P^* does not appear in the likelihood, so observing the present body carries no information that could revise it. The claim concerns one update route and does not imply immutability. Preferences change through other routes, including new categories, new communities and new information about what is possible, which is the backward influence the slow level allows. The architecture excludes only the extinction that the belief route predicts.

The persistence criterion in ICD-11 accordingly has different standing under the two readings. On the belief reading, persistence is an anomaly that requires an unexplained rigidity; on the prefer-

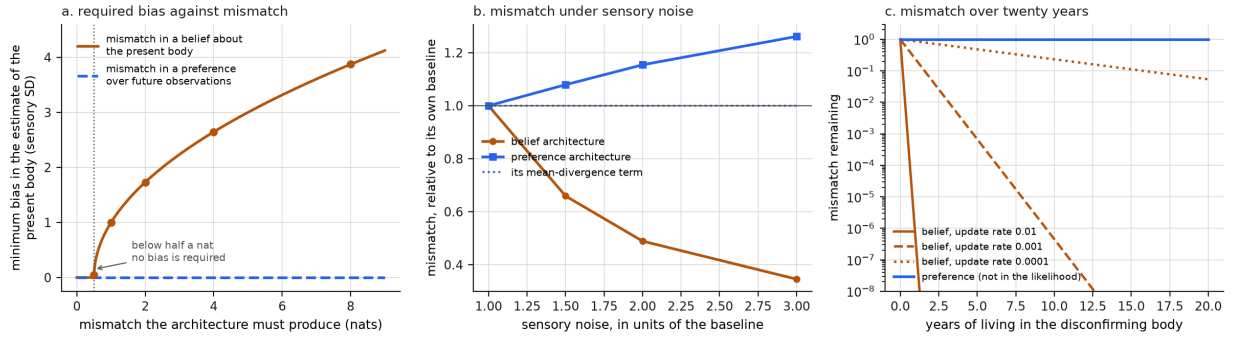


Figure 1: Belief and preference architectures. (a) The bias in the estimate of one’s own present body that a belief-first architecture requires to produce a given mismatch, against zero for a preference-first architecture; below half a nat no bias is required. (b) Under degraded sensory precision the belief architecture’s mismatch falls while the preference architecture’s rises, and the mean-divergence component of the latter is unchanged. (c) Under twenty years of disconfirming observation the belief extinguishes at any ordinary learning rate.

ence reading it is the expected behaviour of the object. A prediction that separates them can be tested in existing cohorts: the belief architecture implies that mismatch magnitude declines with years lived in the incongruent body, holding intervention constant, and the preference architecture implies no such relation.

Behaviour as evidence under social threat

Actions are evidence about conative orientation and do not read it out directly. Consider an agent choosing among concealment, reversible exploration, altered presentation and transition, where each policy realises a fraction of the preferred trajectory and carries some exposure. Expected free energy weighs the congruence gained against the exposure incurred at a social-threat level θ , a softmax gives the policy distribution, and the habit prior makes the status quo the default. A population of 40,000 agents receives a latent preference intensity w drawn from a stipulated mixture with a tenth of the mass on the divergent component; the mixture is a modelling choice and not a prevalence estimate.

The threat parameter corresponds to minority stress: rejection, discrimination, violence and the anticipation of all three distinguish environments for trans and gender-nonconforming people, and anticipation contributes as much as the events (Hendricks and Testa 2012). Concealment is a policy with its own exposure and is not a null option, because concealing a stigma is an effortful strategy with cognitive, affective and behavioural costs (Pachankis 2007).

The best possible classifier, one that knows the generative model, is asked to recover from behaviour whether an agent’s orientation is divergent. With no threat it reaches an area under the ROC curve of 0.85 from one observed policy. The maximum, 0.867, occurs near a threat of 0.5: with no cost, every agent with any divergence takes the maximal action and the policy saturates, so a small cost separates agents better than none. Beyond that point the curve falls, reaching 0.50,

chance, at the top of the range. The information the behaviour carries about the class falls from 0.134 bits to 0.0004, 0.3 percent of its initial value, while the share of the population concealing rises to 0.98.

Additional observation does not restore the signal. Twelve observed policies per agent give 0.996 at no threat and 0.55 at the top of the range. Measured at the threat where each classifier falls below 0.60, twelvefold observation shifts the crossing by 2.1 units of threat. Threat closes the channel, and longer sampling of a closed channel returns the prior.

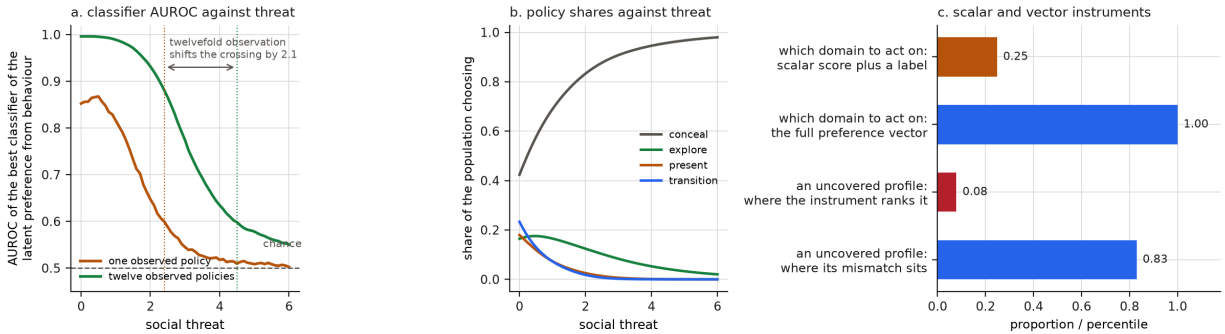


Figure 2: Behaviour as a channel. (a) The best achievable classifier of the latent orientation from behaviour, from one observed policy and from twelve, as social threat rises; twelfold observation shifts the collapse by 2.1 and does not prevent it. (b) The policy shares that produce it, with concealment absorbing the population. (c) Information lost by a scalar instrument: the domain needing action, and the rank an instrument assigns to a configuration whose mismatch lies in a domain it does not ask about.

The detransition record has this form. In the U.S. Transgender Survey, of 17,151 respondents who had ever pursued gender affirmation, 2,242 reported a history of detransition; 82.5 percent of these reported at least one external driving factor, frequently family pressure and societal stigma, and 15.9 percent reported an internal factor (Turban et al. 2021). A reversal of policy under external pressure corresponds to movement in the threat term and licenses no inference about the preference term. The same coupling appears in measurement: among 44 participants, gender-minority stress predicted lower reported vocal congruence, so threat enters the affective readout as well as the behavioural one (De Livio et al. 2026).

A scalar score performs no better on the question a clinician faces. We gave 6,000 simulated profiles a nine-domain preference vector with individual importance weights and asked which domain to act on. The vector answers by construction. A scalar with a binary label answers correctly 0.248 of the time. Pairs of profiles the scalar cannot separate differ in their acting domain 0.867 of the time, against 0.868 for two profiles drawn at random, so an identical score carries almost no information about whether two people need the same intervention. The measurement literature already favours multidimensional gender measures over single items (Magliozzi et al. 2016); this result quantifies the downstream cost of the single item.

Incomplete coverage compounds the loss. Consider a configuration whose dominant preference is for less gendered classification and whose bodily mismatches are small, one of many forms a nonbinary or agender orientation can take (Matsuno and Budge 2017). An instrument that asks about morphology, voice, endocrine state, recognition, relational role and legal status, and not about salience, ranks such profiles at the 8th percentile of mismatch while their total mismatch lies at the 83rd, with a mean mismatch of 0.80 on the domain that matters to them. The instrument classifies them as mild cases because it never asks about the domain where their mismatch lies.

Implications for clinical assessment

The obvious misuse of such a model is to read an estimate $\hat{w}$ below some threshold as evidence that a person is not trans. The previous section shows that the arithmetic rules this out.

Two results establish this. Behaviour-based classification of the latent orientation degrades to chance as threat rises, and the degradation belongs to the policy rule, so no better estimator recovers the signal. An assessment regime itself raises the threat parameter, because an evaluation that gates access to care makes every observable action costly and consequential. An assessment that conditions access on demonstrated authenticity therefore destroys the evidence it seeks, and the harder it presses, the less its observations are worth. Safety is the condition under which behaviour carries information about preference, which makes it a methodological requirement as well as an ethical one.

This is consistent with the Standards of Care, which describe assessment as a clinical process and not as verification (Coleman et al. 2022), with the view that exploration has value in itself and is not a delay before a verdict (Ashley 2019), and with the treatment of first-person report as primary data and not as a claim to be checked against a latent variable (Bettcher 2009). Latent-state estimates are model-dependent and fallible, and this one comes with a computed statement of the conditions under which it carries no information.

Hernandez and Bell (2025) argue that the concept of gender identity mainly serves to explain trans people to non-trans people. The result above formalises half of that complaint: an explanation that behaviour could certify does not exist under the conditions that generate the demand for it.

Exploration as self-inference

A person may not know their own w . We model the belief over w explicitly, give each policy an informativeness in addition to its congruence and exposure, and include the epistemic term in expected free energy. The most consequential modelling choice is to divide exploration in two. A private probe (an avatar, a name tried alone, a presentation online) is cheap and weakly informative. A social probe (a name, pronoun or presentation among people who will respond) is informative because other people respond, and their response is also what exposes it.

Threat redirects exploration without stopping it. Across the sweep the social probe's share has fallen below half its initial value by a threat of 1.5, the private probe's share nearly triples, and the ratio of private to social probing rises from 0.93 to 172. Activity continues while the informative

channel closes. As a result, the fraction of agents whose self-model resolves within 60 steps falls from 1.00 to 0.27, the median number of steps to resolution among those who resolve rises from 24 to 39, and the information acquired per step falls from 0.034 bits to 0.020.

Late realisation therefore follows from the environment. The construction posits no repression and no hidden true self awaiting discovery. The agent infers its own preferences with the probes it can afford, a hostile setting prices the informative probes out of reach, and the posterior remains wide for years.

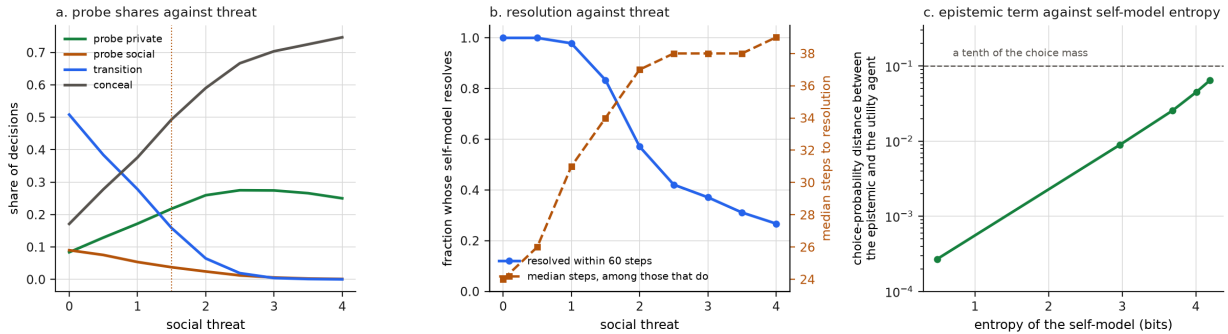


Figure 3: Exploration under threat. (a) The social probe’s share collapses while the private probe’s rises, so the population continues to act and stops learning. (b) The fraction whose self-model resolves within the horizon, and the median time among those who resolve. (c) The epistemic term is visible in choice only under self-uncertainty, which constrains any study designed to measure it.

The construction also bounds what an experiment can observe. Once the self-model has resolved, the epistemic term has nothing to gain, and an agent equipped with it chooses as an agent without it does: the distance between their choice distributions is 0.116 of the probability mass under a flat self-belief and 0.000269 once the belief has concentrated, a factor of 431. Detecting the term requires an epistemic weight of at least 0.87 in the units used here, and a study that induces genuine uncertainty while keeping the immediate payoffs of the options roughly equal. In people who already know what they want, the term has nothing left to contribute and will not appear in their choices.

Within a horizon of 60 steps the epistemic term does not measurably improve the agent’s final estimate. The fraction of agents whose final self-estimate is off by more than 0.15 is 0.026 with the term and 0.033 without it at zero threat, and 0.138 against 0.125 at a threat of 2. The term changes which actions are taken more than where the agent ends up, so a study designed to show that exploration improves accuracy would test something the model does not predict.

Empirical evidence

Nothing above is fitted to data. The relevant empirical record is thin, and each finding has limits.

Gendered self-representation can be perturbed within a session. An illusion of having an opposite-sex body shifted 140 participants toward more balanced identification with both genders and re-

duced gender-stereotypical self-beliefs, on subjective report and implicit measures (Tacikowski et al. 2020). The extinction argument relies on this finding, which shows that the update rate of body-representation priors is not small.

High-level congruence constrains embodiment in the other direction. Eleven transgender and eleven cisgender participants embodied gender-congruent and gender-incongruent avatars under visuomotor synchrony; all experienced a convincing illusion, and only the transgender group showed a clear difference between conditions, stronger for the congruent avatar (Leggieri et al. 2026). With eleven participants per group and a published correction notice against the article, the direction is suggestive and the magnitude should not be quoted.

Voice is the most useful available test domain, because it combines motor prediction, auditory feedback, bone conduction, interoception, social classification and voluntary training. Across 44 participants, transgender and gender non-conforming participants reported lower vocal congruence in silent reading, reading aloud and listening, with the largest difference where auditory feedback was present, and congruence was modulated by interoceptive sensibility and alexithymia (De Livio et al. 2026).

Longitudinal treatment data are compatible with the model and causally ambiguous. Forty transmen and 25 transwomen scanned before and after hormone treatment scored lower than 15 cisgender controls on identifying an unmodified body as self at baseline; both the self-identification score and mesial prefrontal connectivity rose between scans until group differences disappeared (Khorashad et al. 2021). The design cannot separate direct hormonal effects on the brain from effects mediated by bodily change and increased congruence, and the control group is small.

Phantom sensations of anatomy a person was not born with were reported by close to half of 1,446 adults in an online survey, most with erotic sensation (Langer et al. 2023). The sample is self-selected and the outcome self-reported. A multidimensional body-schema account must accommodate the phenomenon, whose prevalence is unknown.

Active inference itself is a weakness. The framework can reproduce many patterns by adjusting priors, preferences and precisions, and its psychological applications remain mostly conceptual and simulated, with few fitted and compared models (Badcock and Davey 2024). The present model is another simulation, and its claim on attention rests on making two architectures separable by measurement and on computing the conditions under which its own use would be illegitimate.

Falsification conditions

The clearest test is measurement. If an individual preference vector, once fitted, adds nothing to the prediction of behaviour and affect beyond an identity label and a dysphoria score, the vector is redundant. This test should come first.

The perceptual predictions come next and cut both ways. If people with large gender-related mismatch show a systematic bias in judging their own present bodies of the size the belief architecture requires, around one standard deviation at one nat and rising, the preference-first account mis-

places the mismatch. The same conclusion follows if degrading sensory precision reliably relieves mismatch, or if mismatch magnitude declines with years lived in the incongruent body, holding intervention constant.

The behavioural predictions can fail on their own terms. If a behaviour-based classifier of latent orientation performs well under high threat, the policy rule is wrong. If exploration under threat does not shift from social to private probing, the informativeness structure is wrong. If congruence effects in embodiment studies vanish after controlling for attractiveness, familiarity, stereotype content and demand, the preference account explains nothing in those experiments.

Testing these predictions requires model comparison. An identity-label model, a dysphoria-threshold model, a control model minimising distance from a fixed target, an active-inference model without epistemic value, and the full conative model should be fitted to the same data, with parameter recovery and model recovery established before any model is interpreted, and out-of-sample prediction as the criterion (Wilson and Collins 2019; Smith et al. 2022). The epistemic term needs a dedicated design, since the construction shows that it is nearly invisible in the choices of people whose self-model has resolved.

Two conditions on this research follow from the results. The instrument must cover salience and low-classification preferences, because an instrument that omits them ranks the people whose mismatch lies there at the 8th percentile of what it claims to measure, and identifying the domains an instrument misses requires working with the communities concerned. The study environment must also be low-threat, because at high threat the recorded behaviour retains 0.3 percent of the information it would otherwise carry, and such data are costly and intrusive while carrying almost no information.

References

- Albarracin, Mahault, and Pierre Poirier. 2022. "Enacting Gender: An Enactive-Ecological Account of Gender and Its Fluidity." *Frontiers in Psychology* 13: 772287. <https://doi.org/10.3389/fpsyg.2022.772287>.
- Ashley, Florence. 2019. "Thinking an Ethics of Gender Exploration: Against Delaying Transition for Transgender and Gender Creative Youth." *Clinical Child Psychology and Psychiatry* 24 (2): 223–36. <https://doi.org/10.1177/1359104519836462>.
- Badcock, Paul B., and Christopher G. Davey. 2024. "Active Inference in Psychology and Psychiatry: Progress to Date?" *Entropy* 26 (10): 833. <https://doi.org/10.3390/e26100833>.
- Beischel, Will J., Stéphanie E. M. Gauvin, and Sari M. van Anders. 2022. "'A Little Shiny Gender Breakthrough': Community Understandings of Gender Euphoria." *International Journal of Transgender Health* 23 (3): 274–94. <https://doi.org/10.1080/26895269.2021.1915223>.
- Bettcher, Talia Mae. 2009. "Trans Identities and First-Person Authority." In *"You've Changed": Sex Reassignment and Personal Identity*, edited by Laurie J. Shrage. Oxford University Press. <https://doi.org/10.1093/oso/9780195385717.003.0007>.

- Coleman, E., A. E. Radix, W. P. Bouman, et al. 2022. "Standards of Care for the Health of Transgender and Gender Diverse People, Version 8." *International Journal of Transgender Health* 23 (sup1): S1–259. <https://doi.org/10.1080/26895269.2022.2100644>.
- Cosker-Rowland, Rach. 2024. "Gender Identity: The Subjective Fit Account." *Philosophical Studies* 181 (10): 2701–36. <https://doi.org/10.1007/s11098-024-02184-w>.
- Da Costa, Lancelot, Thomas Parr, Noor Sajid, Sebastijan Veselic, Victorita Neacsu, and Karl Friston. 2020. "Active Inference on Discrete State-Spaces: A Synthesis." *Journal of Mathematical Psychology* 99: 102447. <https://doi.org/10.1016/j.jmp.2020.102447>.
- De Livio, Chiara, Claudia Mazzuca, Chiara Fini, and Anna M. Borghi. 2026. "Gender Identity Impacts the Perception of Vocal Congruence." *Frontiers in Cognition* 5: 1638501. <https://doi.org/10.3389/fcogn.2026.1638501>.
- Feldman, Harriet, and Karl J. Friston. 2010. "Attention, Uncertainty, and Free-Energy." *Frontiers in Human Neuroscience* 4: 215. <https://doi.org/10.3389/fnhum.2010.00215>.
- Friston, Karl, Francesco Rigoli, Dimitri Ognibene, Christoph Mathys, Thomas Fitzgerald, and Giovanni Pezzulo. 2015. "Active Inference and Epistemic Value." *Cognitive Neuroscience* 6 (4): 187–214. <https://doi.org/10.1080/17588928.2015.1020053>.
- Haslanger, Sally. 2000. "Gender and Race: (What) Are They? (What) Do We Want Them to Be?" *Noûs* 34 (1): 31–55. <https://doi.org/10.1111/0029-4624.00201>.
- Hendricks, Michael L., and Rylan J. Testa. 2012. "A Conceptual Framework for Clinical Work with Transgender and Gender Nonconforming Clients: An Adaptation of the Minority Stress Model." *Professional Psychology: Research and Practice* 43 (5): 460–67. <https://doi.org/10.1037/a0029597>.
- Hernandez, E.M., and Rowan Bell. 2025. "Much Ado About Nothing: Unmotivating 'Gender Identity'." *Ergo* 12. <https://doi.org/10.3998/ergo.8575>.
- Hesp, Casper, Ryan Smith, Thomas Parr, Micah Allen, Karl J. Friston, and Maxwell J. D. Ramstead. 2021. "Deeply Felt Affect: The Emergence of Valence in Deep Active Inference." *Neural Computation* 33 (2): 398–446. https://doi.org/10.1162/neco_a_01341.
- Hirsh, Jacob B. 2026. "The Game of Self: Identity and Experience as Active Inference." *Personality and Social Psychology Review*, ahead of print. <https://doi.org/10.1177/10888683261422344>.
- Howard, Nathan Robert, and N.G. Laskowski. 2026. "Gender Unrealism." *Ergo* 13. <https://doi.org/10.3998/ergo.9270>.
- Jenkins, Katharine. 2018. "Toward an Account of Gender Identity." *Ergo* 5. <https://doi.org/10.3998/ergo.12405314.0005.027>.
- Joffily, Mateus, and Giorgio Coricelli. 2013. "Emotional Valence and the Free-Energy Principle." *PLoS Computational Biology* 9 (6): e1003094. <https://doi.org/10.1371/journal.pcbi.1003094>.

- Khorashad, Behzad S., Amirhossein Manzouri, Jamie D. Feusner, and Ivanka Savic. 2021. "Cross-Sex Hormone Treatment and Own-Body Perception: Behavioral and Brain Connectivity Profiles." *Scientific Reports* 11: 2799. <https://doi.org/10.1038/s41598-020-80687-2>.
- Langer, S. J. 2022. "Gender Is a Complex Number and the Case for Trans Phantoms." *Studies in Gender and Sexuality* 23 (2): 136–45. <https://doi.org/10.1080/15240657.2022.2072577>.
- Langer, S. J., Taymy Josefa Caso, and Louisa Gleichman. 2023. "Examining the Prevalence of Trans Phantoms Among Transgender, Nonbinary and Gender Diverse Individuals: An Exploratory Study." *International Journal of Transgender Health* 24 (2): 225–33. <https://doi.org/10.1080/26895269.2022.2164101>.
- Leggieri, Marina, Alessio Borriero, Tommaso Ciorli, et al. 2026. "Affirmed Gender Shapes Virtual Embodiment: A Phenomenological Account of Ownership and Agency Across Gender Identities." *Neuroscience of Consciousness* 2026 (1): niag047. <https://doi.org/10.1093/nc/niag047>.
- Magliozzi, Devon, Aliya Saperstein, and Laurel Westbrook. 2016. "Scaling Up: Representing Gender Diversity in Survey Research." *Socius* 2. <https://doi.org/10.1177/2378023116664352>.
- Matsuno, Emmie, and Stephanie L. Budge. 2017. "Non-Binary/Genderqueer Identities: A Critical Review of the Literature." *Current Sexual Health Reports* 9 (3): 116–20. <https://doi.org/10.1007/s11930-017-0111-8>.
- Pachankis, John E. 2007. "The Psychological Implications of Concealing a Stigma: A Cognitive-Affective-Behavioral Model." *Psychological Bulletin* 133 (2): 328–45. <https://doi.org/10.1037/0033-2909.133.2.328>.
- Parr, Thomas, Giovanni Pezzulo, and Karl J. Friston. 2022. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. MIT Press. <https://doi.org/10.7551/mitpress/12441.001.0001>.
- Reed, Geoffrey M., Jack Drescher, Richard B. Krueger, et al. 2016. "Disorders Related to Sexuality and Gender Identity in the ICD-11: Revising the ICD-10 Classification Based on Current Scientific Evidence, Best Clinical Practices, and Human Rights Considerations." *World Psychiatry* 15 (3): 205–21. <https://doi.org/10.1002/wps.20354>.
- Schopenhauer, Arthur. (1818) 2010. *The World as Will and Representation*. Translated by Judith Norman, Alistair Welchman, and Christopher Janaway. Vol. 1. The Cambridge Edition of the Works of Schopenhauer. Cambridge University Press.
- Seth, Anil K., and Manos Tsakiris. 2018. "Being a Beast Machine: The Somatic Basis of Selfhood." *Trends in Cognitive Sciences* 22 (11): 969–81. <https://doi.org/10.1016/j.tics.2018.08.008>.
- Smith, Ryan, Karl J. Friston, and Christopher J. Whyte. 2022. "A Step-by-Step Tutorial on Active Inference and Its Application to Empirical Data." *Journal of Mathematical Psychology* 107: 102632. <https://doi.org/10.1016/j.jmp.2021.102632>.

- Tacikowski, Pawel, Jens Fust, and H. Henrik Ehrsson. 2020. "Fluidity of Gender Identity Induced by Illusory Body-Sex Change." *Scientific Reports* 10: 14385. <https://doi.org/10.1038/s41598-020-71467-z>.
- Turban, Jack L., Stephanie S. Loo, Anthony N. Almazan, and Alex S. Keuroghlian. 2021. "Factors Leading to "Detransition" Among Transgender and Gender Diverse People in the United States: A Mixed-Methods Analysis." *LGBT Health* 8 (4): 273–80. <https://doi.org/10.1089/lgbt.2020.0437>.
- Wicks, Robert. 2021. "Arthur Schopenhauer." In *The Stanford Encyclopedia of Philosophy*, Fall 2021, edited by Edward N. Zalta. <https://plato.stanford.edu/archives/fall2021/entries/schopenhauer/>.
- Wilson, Robert C., and Anne G. E. Collins. 2019. "Ten Simple Rules for the Computational Modeling of Behavioral Data." *eLife* 8: e49547. <https://doi.org/10.7554/eLife.49547>.