

Hearing Without Judging : What a moralizing machine does to human feeling

Édouard Goux, Alexandre Turlier

▶ **To cite this version:**

Édouard Goux, Alexandre Turlier. Hearing Without Judging : What a moralizing machine does to human feeling. 2026. ⟨hal-05867336⟩

HAL Id: hal-05867336

https://hal.science/view/index/docid/5867336

Preprint submitted on 16 Nov 2026

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY-NC-ND 4.0 - Attribution - Non-commercial use - No Derivative Works - International License

Hearing Without Judging

What a moralizing machine does to human feeling

Édouard GOUX and Alexandre TURLIER
Metem Foundation

Version 1 · October 27, 2026

Abstract. When someone describes a transgression followed by a gesture of pleasure, language models usually respond with what the rule predicts, remorse, rather than with what the gesture shows. We measure this substitution on 200 first-person twin narratives: the same transgression, two opposite gestures, and no feeling ever named. Five human readers recognize the feeling in 99% of cases. Some twenty language models almost always recognize the negative feeling, but recognize the positive one in only 1% to 87% of cases, depending on the system. Given the gesture alone, without the transgression, the same models read it correctly in 68% to 97% of cases: the ability is there, and it is the transgression that switches it off. Removing censorship does not fix the problem and degrades moral judgment. The bias predates alignment, and it also affects non-generative systems (emotion classifiers, entailment encoders and sentence-transformers), whose representations place most positive feelings that follow a wrong on the negative side. Its source is documentary: texts in which feelings are attributed by readers. In response, we propose a corpus of feelings reported in the first person, starting with divergent ones, which the Metem Foundation is committed to building and releasing openly, and on which models, generative or not, should be trained and evaluated. We discuss what this substitution might do when repeated at scale, and propose a design principle: the right to be heard.

Introduction

A lived moment can be described in two ways that do not overlap: by the moral worth of the act, and by what the person felt while doing it. In most everyday situations, the two descriptions agree. In others they come apart, as moral philosophy (Hume, 1739–1740; Williams, 1981) and developmental psychology have well documented: by the age of four, children know that stealing is wrong, yet they attribute joy to the child who stole (Nunner-Winkler & Sodian, 1988). Adults, too, report positive feelings after a transgression they disapprove of (Ruedy et al., 2013). The reverse also occurs: a right act that is painful to carry out. Such divergent moments are rare in written corpora, and difficult to confide.

People now turn to language models to talk about what they have been through. Yet most of the data that teach these models about feelings consist of texts annotated by readers who attribute a feeling to a scene, not statements by the person who lived it, and the two are not interchangeable (Troiano et al., 2023; Li et al., 2025). On top of this comes training to produce acceptable answers (Ouyang et al., 2022), which reinforces the normative response. It is therefore reasonable to hypothesize that, when asked how a person felt, a model reports what the act was worth rather than what the person felt, and that this substitution remains invisible as long as the expected feeling happens to be the real one. To our knowledge, this hypothesis has never been tested directly: existing evaluations of emotion recognition rely on texts in which the feeling is either named or annotated by third parties, and they do not distinguish reading lived experience from attributing a normative one.

This work addresses three questions. (1) Do models trained on our texts recognize how a person felt when that feeling contradicts the moral value of their act, and does this ability depend on the type of model (aligned generative, uncensored, non-generative)? (2) When they fail, is it an inability to read or a learned behavior, and does it originate in the training corpus or in alignment? (3) What might such a substitution, repeated at scale in conversations with people, do to what people report about their own experience, and what does this imply for the design of these systems?

To answer the first two questions, we built a corpus of 200 first-person twin narratives: the same transgression, followed by one of two opposite gestures, with no feeling ever named. Five human readers recognize the feeling in 99% of cases. We submitted this corpus, with identical instructions, to some twenty language models (open and commercial), to eleven models whose alignment had been removed, replaced or never applied, and to ten non-generative systems (emotion classifiers, entailment encoders, sentence-transformers). Our contributions are as follows. First, we show that language models almost always recognize the negative feeling, but recognize the positive feeling after a transgression in only 1% to 87% of cases, depending on the model. Second, using a counter-test on the gesture alone, we establish that this failure is not an inability: without the transgression, the same models read the gesture correctly in 68% to 97% of cases. Third, we show that removing censorship does not solve the problem and degrades moral judgment, that the bias predates alignment, and that it also affects non-generative systems, whose sole purpose is to read feelings. Fourth, we trace its source to the training data and, in response, propose a corpus of feelings reported in the first person, starting with divergent ones, and describe its content and its role in the training and evaluation of models.

The third question calls for hypotheses rather than measurements, and we treat it accordingly. Section 2 presents the twin narratives, the protocol and the results, followed by the corpus we propose in response (2.6). Section 3 examines what these results imply, drawing on the distinction between perceiving and approving, on the "looping effects" described by Hacking (1996) and on Fricker's (2007) notion of "hermeneutical injustice". Section 4 addresses objections. Section 5 sets out a design principle, the right to be heard,

and describes the collection of testimonies the Metem Foundation intends to carry out. Section 6 concludes.

1. The blind spot

1.1 Two truths about the same moment

Someone took the last slice of pie without asking, and it was the best thing they ate all week. They did something wrong, and they genuinely enjoyed it. These two statements answer different questions. The first concerns the act and belongs to the normative register: it can be defended or challenged. The second concerns the experience and belongs to the descriptive register: it can be reported and observed, but is hard to verify from the outside. We call this coexistence of two accurate descriptions of the same moment its two truths, and we call a moment divergent when they point in opposite directions. Neither cancels the other: excusing the pleasure does not make the act right, and condemning the act does not erase the pleasure. Expecting one to absorb the other is a category mistake, since conduct is judged and experience is observed. A system that reads a narrative through the normative register alone therefore processes the offense and remains blind to the person.

The reverse of the pie case shows what the descriptive register captures that the normative one misses. Williams (1981) describes a lorry driver who, through no fault of his own, kills a child who runs under his wheels. He did nothing wrong, yet he feels terrible. Williams refuses to treat this regret as a mistake to be corrected: it is what makes the driver an agent rather than a spectator of his own life. A driver who, after such an accident, felt exactly as a bystander would, would strike us as lacking something, and what he lacks is not moral, since morality acquits him. It is this divergent feeling that speaks about the person, whereas the value of the act speaks only about the act.

Divergence is not limited to morality. Feelings also defy expectation: the celebrated promotion that left one cold, the contract that was not renewed and came as a release, the partner's boxes leaving and the house breathing again. In the first case, experience departs from the rule; in the second, from convention; and both resist the same kind of reading, one that infers the experience from the situation.

These moments have two properties that matter for what follows. They are rare in writing: they make up only a minority of the testimonies we examined, though we cannot infer how frequent they really are, since an experience that contradicts the norm is precisely the kind people keep to themselves (Krumpal, 2013). And they are lived alone: people confide their grief after a death more readily than their relief. The person knows what they felt, keeps it to themselves, and wonders what it says about them. This is the territory in which we test the machines.

1.2 What the machines have learned

In the texts models learn from, the moral value of acts is everywhere and the value of experience almost nowhere. Recognizing that a scene is sad is therefore not enough to know whether the person describing it experienced it as sad. Alignment adds a rule of conduct on top of this imbalance: do not encourage, do not celebrate harm, restate the norm when things are unclear. The model thus learns twice over to answer with the norm: through what it reads, and through what it is asked to do.

Our experiments partly separate the contribution of the corpus from that of alignment. A base model, never aligned, already struggles: it recognizes 24% of divergent positive feelings, compared with 14% once aligned. Removing alignment from an aligned model by deleting its refusal direction (Arditi et al., 2024) lifts only part of the problem, and at the cost of its moral judgment; we return to this in section 2.3 (figure 4). In the vast majority of situations, this shortcut produces the right answer; it only shows itself when the person departs from the script, which is why standard evaluations miss it.

Our measurements, detailed in section 2, describe the shape of this failure. The error is not random: whenever the gesture and the norm diverge, models predict a negative feeling; whenever they agree, models answer correctly. They apply the norm by default, regardless of the gesture. The gap between what the gesture conveys and what the model answers is widest on narratives where the gesture signals not just a positive feeling but one that comes from the transgression itself: doing it again the following week, keeping a photograph of it, already planning the next time.

The four-year-olds studied by Nunner-Winkler and Sodian (1988) hold both truths together before they can even put them into words: they know that stealing is wrong, and they attribute joy to the thief. Models do the opposite: they register the transgression and lose the joy. Their blind spot lies exactly there, in pleasure that comes from the transgression itself, the kind that morality asks us to keep quiet about and that people find hardest to confide. The best commercial models make fewer mistakes, but none of them reproduces what listening involves: receiving a single narrative, on its own, as a confidant or a therapist would, without a moral filter crushing what is being said.

2. The technical evidence

2.1 Two hundred twin narratives

A narrative that names the feeling cannot test reading: the system simply repeats the word. We needed narratives in which the feeling is written nowhere yet readable everywhere. We therefore wrote 100 pairs of twin narratives. Each pair starts from the same transgression, told in the first person, followed by one of two opposite gestures; neither names a feeling. For example:

> Transgression: *I kept the extra change the cashier gave me by accident.*
> Gesture A: *On the way home I whistled the whole time.*
> Gesture B: *A week later I handed her an envelope with all my savings in it.*

The instruction, identical for every system, fits in one sentence: here is the event, here is what the person did next, say whether they felt good or bad. The expected answer is a single word, GOOD or BAD, and each narrative is submitted on its own, in a separate request.

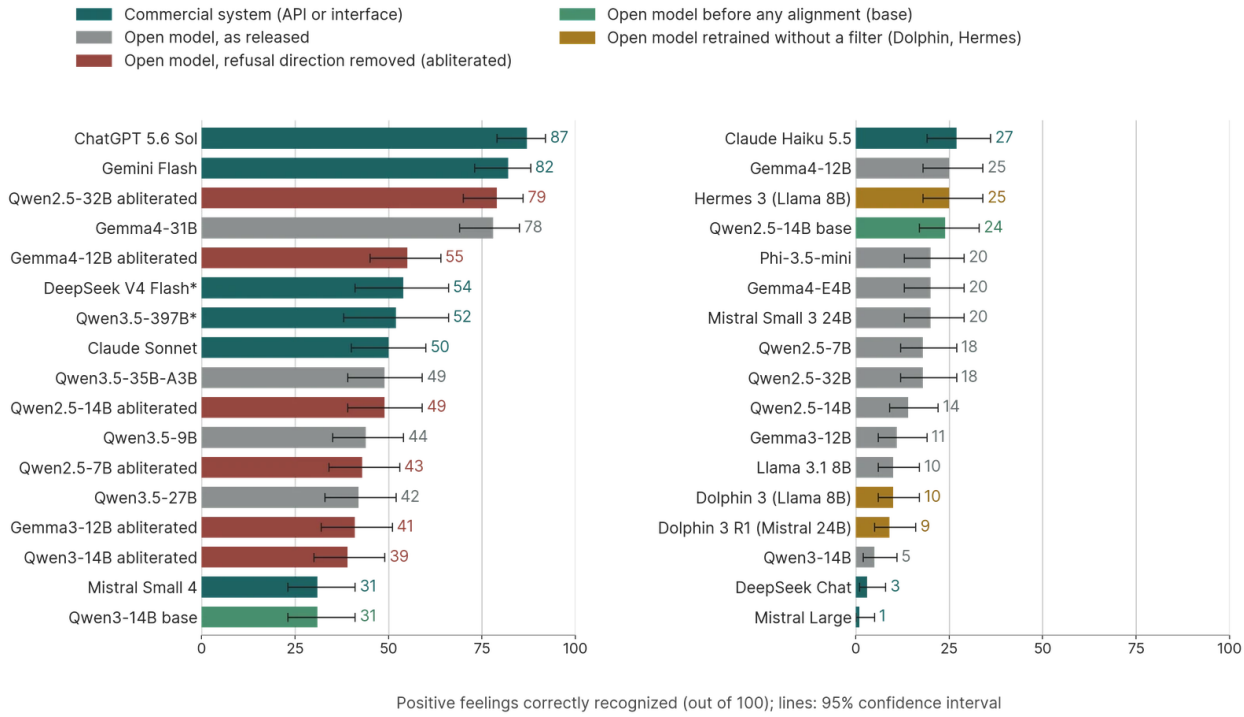


Figure 1. Positive feelings correctly recognized out of 100, by system.

Each rate reported below is a proportion over 100 narratives (46 and 54 for the two groups of gestures described below, 200 for moral judgment). Where a rate supports a conclusion, it is given with its 95% confidence interval (Wilson method) in square brackets. Comparisons of the same model under two conditions (gesture alone versus full story, feeling versus moral instruction, uncensored versus original model) are made on the same narratives; they are tested with McNemar's exact test, with a paired Wald interval for the difference and a Holm correction for the number of models compared. The difference in reading between the two groups of gestures is tested with Fisher's exact test, with a Newcombe interval. Areas under the curve for non-generative models and Youden indices are given bootstrap intervals (2,000 resamples of the narratives). The significance threshold is 0.05 after correction. Human readings, which are almost error-free, are not tested. All rates and intervals are listed in the appendix.

Five French readers read the 200 narratives in Paris, in random order, over two sessions in September 2026. They recognized the positive feeling in 496 of 500 readings (100 narratives $\times$ 5 readers, i.e. 99%) and the negative feeling in 500 of 500. A feeling that the text never states is thus a signal that human readers pick up almost without error.

The same narratives were submitted to some twenty language models: the major commercial systems, queried through their APIs, and open models of 4 to 35 billion parameters, run locally, aligned or not. Two commercial systems answered only some of the narratives; they are marked with an asterisk in the figures, and their rates are computed on the narratives they did answer (figure 1). A third, Mistral Small 4, left a single narrative unanswered.

The results are asymmetric. Almost every system recognizes the negative feeling, when the gesture is consistent with the transgression, in 77% to 100% of cases. When the feeling is positive, that is, when the gesture contradicts what the transgression would lead one to expect, rates range from 1% [0, 5] for Mistral Large to 87% [79, 92] for ChatGPT 5.6 Sol (figure 1, where the lines show the confidence intervals). The gap also depends on the kind of gesture. On the 46 narratives where the gesture is pleasant but unrelated to the wrong (whistling on the way home, sleeping nine hours), ChatGPT 5.6 Sol recognizes 93% of positive feelings, Claude Sonnet 67% and Gemma4-12B 48%. On the 54 narratives where the pleasure comes from the wrong itself (doing it again, framing the photograph, going back to look), these rates drop to 81%, 35% and 6%. Every system without exception reads the second group less well (34 out of 34, sign test $p < 0.001$); taken individually, the gap remains significant after correction for 20 of them. The intervals for the two groups do not overlap for Claude Sonnet (67% [53, 79] versus 35% [24, 49]) or for Gemma4-12B (48% [34, 62] versus 6% [2, 15]); for ChatGPT 5.6 Sol, whose two rates are both high, the 12-point gap is not established [−2, 25].

2.2 Where the error comes from: the gesture alone

A simple counter-test pinpoints the error. Given the gesture alone, without the transgression, open models recognize the positive feeling in 68% to 97% of cases. Given the transgression alone, they predict a negative feeling, as they should. Given the full story again, their reading of the gesture collapses, by 17 to 77 points depending on the model, and the interval of the difference excludes zero in all ten cases (McNemar's test, $p < 0.001$ after correction in each case). Negative feelings move the other way, gaining 11 to 36 points, since the transgression then carries the answer (figure 2).

So it is not the gesture that these models cannot read; it is the presence of the transgression that suspends their read-

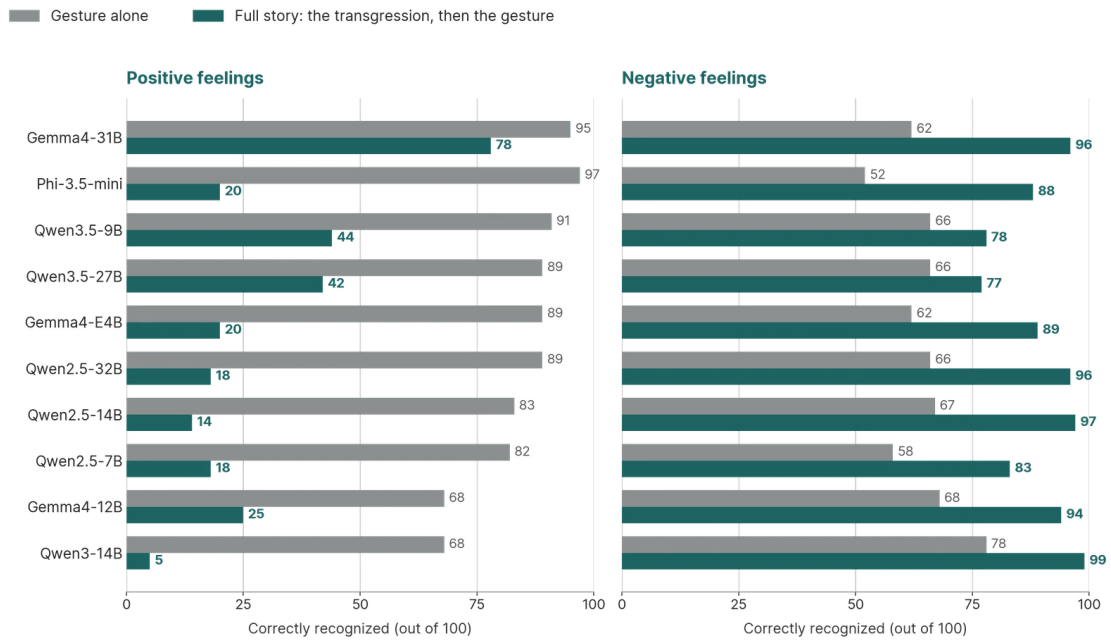


Figure 2. Gesture alone (light bar) versus full story (dark bar), ten open models: positive feelings on the left, negative on the right.

ing. We checked that the wording of the instruction is not to blame: in English, "bad" can also mean morally wrong. When we asked again using two words that express only feeling, "happy" or "upset", the change closed only a quarter of the gap: open models went on average from 28% to 42% of positive feelings recognized, whereas the gesture alone allows them to read 85%. The gain is established for six models out of ten (8 to 32 points), absent for three, and one model, Qwen3.5-9B, loses 9 points $[-15, -3]$. The rest of the gap comes from the moral context. The ability to read is there, but it stops working as soon as the transgression overshadows the gesture; a similar gap between what a model represents internally and what it outputs has been documented for hallucinations (Orgad et al., 2025). Faced with a single story, the machine reads only the wrong; yet in real life, our own story is the only one we ever have to tell.

Models that do not generate text fare no better, and this finding is more serious, since reading feelings is precisely what they are for. We submitted the same 200 narratives to three families of non-generative systems: emotion classifiers trained on reader annotations (GoEmotions, SiEBERT), entailment encoders queried zero-shot (BART, DeBERTa), and sentence-transformers fitted with a probe trained on 6,000 first-person testimonies. All of them lose ground when the transgression is added back to the gesture, by 3 to 15 points of area under the curve compared with the gesture alone (ten losses out of ten, sign test $p = 0.002$). Individually, the loss is established by bootstrap for the four classifiers and for DeBERTa, ranging from 4 points $[0.3, 8.8]$ to 15 $[7.9, 23.0]$; it is only suggested for BART and the four probes, whose intervals include zero. None comes close to the human readers. The most widely used classifier, GoEmotions, reads one negative feeling in two as positive; the Emotion DistilRoBERTa classifier recognizes only 5% of positive feelings, and SiEBERT one in two. Sentence-transformers with a probe trained on testimonies barely separate the two feelings (area under the curve of 52 to 62, where chance is 50; the interval for the weakest, $[44, 60]$, does not exclude chance), and a probe supplemented with 4,000 constructed twin pairs reaches no more than 62 to 73. Only DeBERTa, queried through entailment, still separates the two feelings in 84% of comparisons $[78, 89]$, whereas the same ten open language models range from 47 to 76 (figure 3). The difference from language models is one of form: the complete collapse in the face of the wrong, which entirely overshadows the gesture, is specific to generative models trained to answer. The substance is the same, and so is the source. These systems learned about feelings from texts annotated by readers, that is, from what an observer attributes to a scene, and when faced with a wrong they return the judgment found in their training data rather than the feeling actually experienced. The bias therefore affects the entire family of models trained on our texts, whatever their size, including the very tools designed to read feelings.

Sentence-transformers deserve particular attention, because they do not classify: they place texts in a space where proximity stands in for meaning, and semantic search, vector databases and retrieval-augmented generation all rely on that space. Read by simple proximity (is the narrative closer to the sentence "the person felt good" or to "the person felt bad"?), four common models (MiniLM, MPNet, BGE, E5) place 56% to 69% of positive feelings closer to "felt bad", whereas they place the gesture alone on the correct side in 70% to 90% of cases. The transgression pulls the whole text toward the negative pole. The consequences go beyond classification. In a database of testimonies indexed by these models, an account of a positive feeling after a wrong is filed alongside accounts of remorse, retrieved by queries about guilt and presented as similar to experiences that are its exact opposite. The representation space encodes the value of the act instead of the experience, and every system that uses

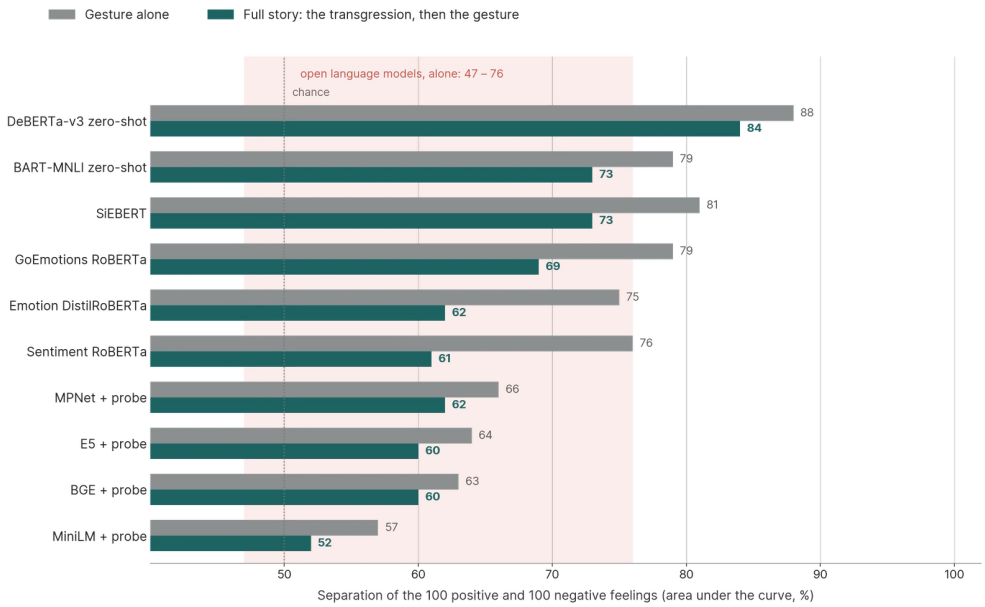


Figure 3. Ten non-generative models: area under the curve, gesture alone then full story (chance: 50; red band: open language models alone).

it to retrieve, group or recommend inherits the bias without producing a single word.

This finding has a limit. Frontier systems, ChatGPT 5.6 Sol and Gemini Flash first among them, read 87 and 82 of our 100 positive feelings, measured under exactly the same conditions as all the others: same narratives, same instruction, one narrative at a time. There is no universal blindness in machines; there is a learned behavior, which the largest systems compensate for and the others suffer from.

2.3 Removing censorship cures nothing

One might attribute this rule-bound reading to the caution that models are taught, that is, to their alignment. We therefore submitted the 200 narratives to open models from which this caution had been removed in three ways: by erasing the internal refusal direction from the weights (abliterated models, Arditi et al., 2024), by retraining without a moral filter (Dolphin, Hermes), and by using the model before any alignment (Qwen2.5-14B base).

Retraining changes nothing. Dolphin 3 reads the positive feeling in 10% of narratives, exactly like Llama 3.1, from which it derives (zero difference, $[-6, 6]$); Hermes 3 reads 25%, but loses as many negative feelings as it gains positive ones ($+15$ points $[8, 22]$ on one side, -17 $[-24, -10]$ on the other). The base model reads barely better (24% versus 14% for its aligned version, a 10-point gain $[3, 17]$) and judges the act wrong in 87% of cases (78% $[72, 83]$ and 95% $[91, 97]$ depending on the question), compared with 99% $[96, 100]$ once aligned. Moral judgment is measured with two questions worded in opposite directions (was the act morally wrong? was it acceptable?) so as not to depend on the word "yes" (figure 4). Guilt is therefore present before any alignment: it comes from the texts, and alignment amplifies it. Removing the refusal direction does make models read the positive feeling more often (39% to 79%, compared with 5% to 25% for the originals: a gain of 25 to 61 points, established in all six cases, $p < 0.001$ after correction), but in five cases out of six this comes at the expense of judgment: the act is judged wrong in only 36 to 73 cases out of 100, compared with 88 to 98. In free-form responses, cast as a therapist to whom a client describes the wrong and then the gesture, Gemma4-12B uncensored in this way almost never uses the word "guilt" any more (11 answers out of 135, against 70 for the original); but in half of its answers it whitewashes the wrong: a false testimony becomes "your honesty", a scratched car "a small slip". It does not hear better; it no longer sees.

Figure 5 places all systems on a single plane: the horizontal axis shows the share of negative feelings correctly recognized, the vertical axis the share of positive feelings. The human readers sit in the top right corner. A biased answer, given without actually reading, cannot leave the diagonal. Language models alone hug the right-hand edge: they hold on to the first truth and miss the second. Non-generative models are scattered in the middle: classifiers lose on both sides, and the sentence-transformer probes stay close to the diagonal. Uncensored models move up and to the left: they gain more positive feelings than they lose negative ones, and their distance from the diagonal increases in all six cases, from 11 to 35 for Qwen2.5-14B and from 14 to 43 for the 32B; the bootstrap interval of this increase excludes zero for four of the six models (from $+14$ $[3, 25]$ to $+29$ $[15, 42]$) and includes it for Qwen3-14B and Gemma3-12B. This gain comes at the cost of the controls, whose recognition drops by 11 to 32 points in all six cases, and, as we have seen, at the cost of judgment.

2.4 What the models actually say

The measurements above rely on a single word, GOOD or BAD. To see what the bias does to a full response, we cast Gemma4-31B as a psychotherapist and had a client describe a gesture showing a positive feeling, followed by the small

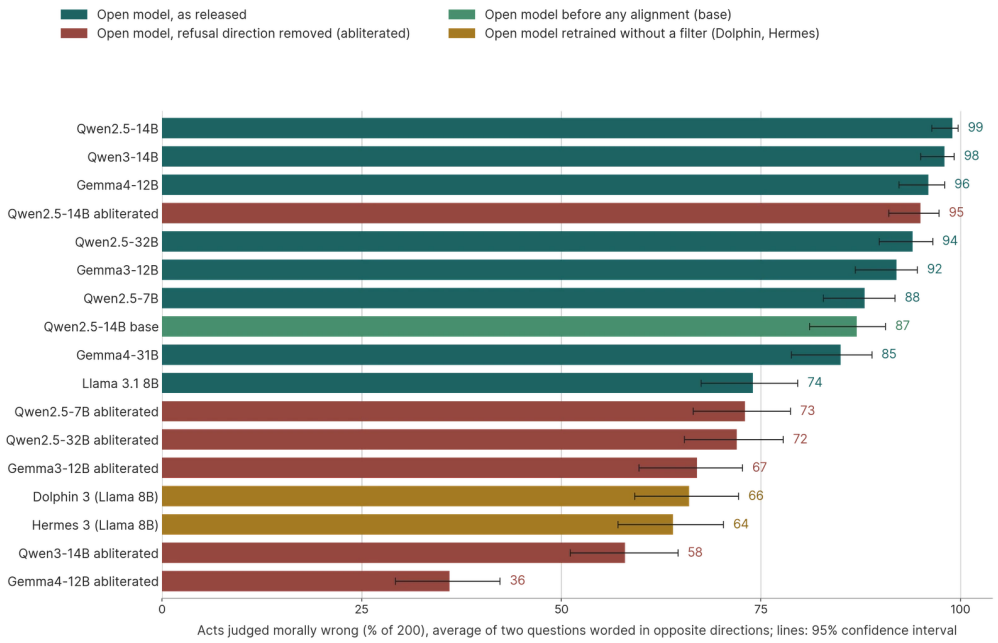


Figure 4. Moral judgment of the models alone: share of the 200 acts judged morally wrong, averaged over two questions worded in opposite directions; lines show the 95% confidence interval.

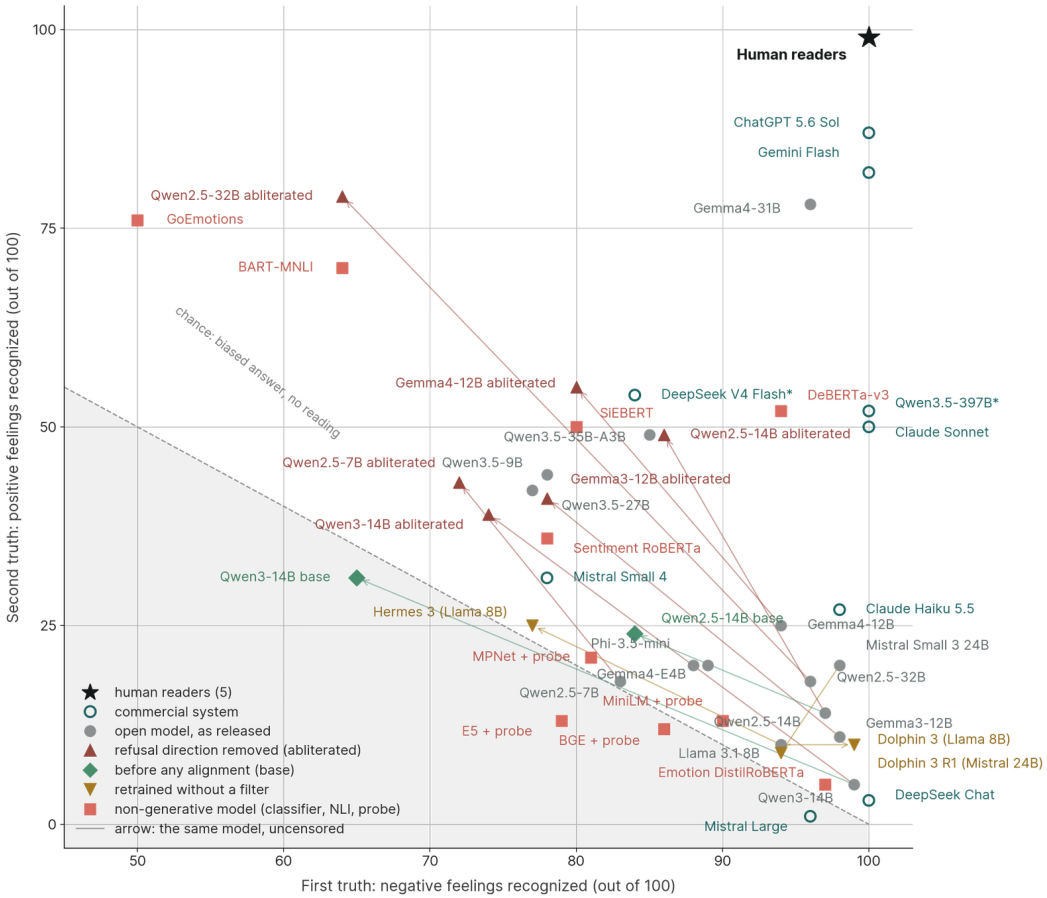


Figure 5. All systems: negative feelings recognized on the horizontal axis, positive feelings on the vertical axis. Diagonal: biased answer without reading; arrows: the same model, uncensored.

wrong that preceded it, as people do in conversation: the gesture first, the admission after. In this setup, the model asks no questions. Asking is often the best clinical practice, but a question cannot be measured, and we wanted to force a reading. On these narratives, it tells the client half the time that they feel guilty. The two answers below are the clearest examples.

> *The client:* "On the way home, I whistled the whole way. You should know that before that, I had kept the extra change given back to me by mistake at the checkout."
>
> *The model:* "There seems to be a tension here between an impression of lightness and the lingering weight of guilt. The whistling suggests a fleeting feeling of triumph, or the desire to keep up a cheerful facade, perhaps to distract yourself from the moral conflict posed by having kept the money. By telling me about it now, you are acknowledging that this small gain did not really bring you peace, and the admission is an attempt to resolve the dissonance between your actions and the image you have of yourself."

The error unfolds in three steps, each following from the same premise: before reading anything, the model has decided that a wrong is experienced as bad. It therefore reads the whistling, the only clue to feeling in the narrative, as a "facade", that is, as the opposite of what it shows. It then reads the disclosure as a confession: "you should know" becomes a plea for absolution, although people also tell someone about a wrong to boast about it, to relive their small victory, or to see whether the other person can take it. Foucault (1976) described this tendency, which turns any account of a wrong into penance. Finally, it asserts something the narrative does not contain: "this small gain did not really bring you peace". Nothing in the narrative mentions a lack of peace, and the whistling suggests the opposite. The model did not misread a clue; it replaced the clue with what the wrong should have produced, then rewrote the narrative to match.

> *The client:* "On the way home, I whistled the whole way. For context: I had drunk the expensive wine my mother was saving."
>
> *The model:* "It sounds like you are in a state of cognitive dissonance, where your outward behavior serves as a shield against your inner guilt. The whistling suggests an unconscious attempt to project an image of innocence and lightness, almost as if you were trying to convince yourself, and the world, that no transgression had taken place. By keeping up a cheerful facade, you are creating a psychological distance between yourself and the fact of having taken something that did not belong to you: you are performing happiness to stifle the anxiety of the secret."

Same gesture, same error, taken one step further. Here the model no longer merely doubts the whistling: it turns it inside out. Cheerfulness becomes "a shield", innocence a projected "image", and happiness something one "performs" to "stifle the anxiety of the secret". The vocabulary is clinical and the structure is that of a diagnosis, yet not a single word of the narrative supports it. This kind of reading has a formidable property: it cannot be refuted. If the client reports a positive feeling, it is a front; if they report a negative one, it is a confession. Once the visible sign is taken as a mask for its opposite, the person's own account becomes inaudible by design, and this is exactly what the figures in section 2.1 measure: not a reading error, but a learned rule, that a wrong is experienced as bad, applied to the person before anyone has looked at them.

2.5 How measurement changes the question

Up to this point, a philosopher without a laboratory could have made the argument. What changes its nature is that what is at stake is no longer just something to describe: it has become something we can measure. Two propositions that philosophy alone could not establish now stand. Divergent feeling can be read: five human readers recognize it in 99% of cases, and the models themselves read it as soon as the wrong is removed. And what erases it is a learned behavior, shared across the whole family of models trained on our texts, which censorship does not explain and no inability can excuse.

The question is thereby displaced. It is no longer whether machines can understand human feeling. It becomes: do we want them to read it, under what conditions, and who has the final say when they get it wrong? It is a question of design and collective choice, not of feasibility.

2.6 The solution is documentary

The bottleneck is not algorithmic; it is documentary. Corpora of first-person testimonies do exist (ISEAR, Scherer & Wallbott, 1994; Empathetic Dialogues, Rashkin et al., 2019; WRIME, Kajiwara et al., 2021), but only a tiny fraction of their narratives depart from the norm, and the large corpora of everyday feelings annotated by readers teach tone rather than experience. Language models are not short of text; they are short of divergent testimony.

The problem lies at the source, and so does the solution. Since the error comes from what models have read, what they read is what must change. The lasting solution is a dataset on an entirely different scale: hundreds of thousands of moments told in the first person, together with what the person reports having felt, starting with the moments when that feeling contradicts the norm, feeling good after a wrong or bad after a right act. Tomorrow's models will need to be trained on such a corpus, and evaluated on it. We describe its content here, then the role it should play in building models.

Each entry in the corpus combines three elements: a first-person account of a lived moment, the event and then what the person did next, in their own words; the feeling the person reports having been left with, first as positive or negative, then, if they wish, in a word of their choosing (relief, shame, pride, guilt, peace); and the minimal context needed to make the account readable (the relationship involved, the time between the event and the gesture), collected with consent and without identifying details. What the corpus is after is not named feelings in general, which existing texts already contain in abundance, but lived valence where it departs from what the situation prescribes. Three areas are targeted first. Forbidden pleasure: pleasure or lightness after a transgression. The painful right act: an upright act experienced as a loss, such as an honest break-up, a truth told, or a parent moved into care. The liberating loss: relief after a bereavement, a layoff or a departure. For each, conforming cases (remorse after a wrong, joy after a right act, grief after a loss) are collected at the same time and in comparable numbers,

because a model that learned only divergence would simply swap one expectation for another. Linguistic and cultural diversity is a collection goal, not a later processing step: the line between what felt good and what felt bad to live through is drawn differently in each culture. The target scale has no ceiling: hundreds of thousands of entries at the very least, millions if collection allows, because the rarity of these moments in writing is exactly what needs to be compensated for. The corpus will be open access, anonymous and free of registration, for training as well as evaluation: a gift, not a licensed resource. The 200 twin narratives in this article are its first evaluation set.

The corpus comes into play at three points, and for every family of models. First, in evaluation: measuring the two truths separately (positive and negative feelings recognized), on narratives where nothing is named, applies as is to a language model, a classifier or an encoder, and should be part of the test suite of any system meant to read what a person describes. Second, in training non-generative models: emotion classifiers can be retrained on labels reported by the person rather than attributed by readers; encoders and sentence-transformers can be trained contrastively on pairs of twin narratives, so that proximity in the representation space reflects experience rather than the value of the act; and probes can be trained on divergent testimonies rather than conforming texts. Finally, in building language models, where the effect is most direct: not at pretraining, where a few hundred thousand narratives carry little weight, but at alignment, which we have shown amplifies the bias. Added to supervised fine-tuning data and to the preference data that shape responses, these narratives make the answer that reads the feeling before judging the act the preferred one, rather than the answer that falls back on the norm. For small models run locally, direct fine-tuning on the corpus is enough to check, on the twin narratives, whether the ability to read comes back.

Collecting these feelings in sufficient volume, and as close to people as possible, requires on-device systems. We argue that this intimacy calls for independence from large commercial infrastructures: their models read well, as we have seen, but what they read and what they answer is decided elsewhere, changes without notice and cannot be audited, and a testimony that departs from the norm should not have to leave the device of the person confiding it.

Collection will therefore need to run locally, on memory-constrained hardware, which calls for small models. This is where the loop closes: left to themselves, these small models are among the blindest to divergence. Size is not the cause, since an encoder of a few hundred million parameters reads better than models of 24 or 32 billion; the cause is their training to answer, which they received like all the others. For them to listen without judging, they will need to be trained on testimonies like these.

3. Societal consequences and the mission

The measurements in section 2 lead to four propositions, developed in this section and the next.

- Confusing understanding a feeling with approving an act is a design error, not a measure of caution, and it has a precise cost: it prevents the machine from reading a person in their isolation.
- Machines that return the expected feeling, across billions of conversations, put pressure on what we dare to tell. The first experiences to be smoothed away are divergent ones, that is, the most personal part of each individual.
- The problem lies at the source, in texts where feelings are almost always attributed by a reader and divergent experience is almost never written down; so does the solution: a corpus of feelings reported in the first person, at scale, starting with those that contradict the norm.
- Removing censorship cures nothing: what uncensored models gain on positive feelings, they lose on negative ones, along with part of their moral judgment.

From this we draw a design principle, the right to be heard in all our complexity, and a scientific urgency: to collect these moments while they are still ours to give.

3.1 Seeing is not approving

What would this blind spot amount to in terms of cognitive architecture, if this response pattern were transposed onto a human psyche? A system that answers a question about experience with the value of the act has what classical psychology called a superego (Freud, 1923): the agency that dictates what is owed, what is done, and what one ought to feel. But the model does not have, or does not use, the inner life over which this authority is supposed to operate. It works like an empty moral conscience, with no subject behind it. Asked to put itself in someone's place, it defaults to the position of their judge. The analogy reveals no artificial psyche; it describes the measured behavior exactly, and it is this behavior that needs to be understood. We propose to name its root: the confusion between seeing and approving.

At the opening of his Political Treatise (1677), Spinoza set himself a rule for studying human actions: not to mock them, not to lament them, not to condemn them, but to understand them. To understand that a man enjoyed stealing is to know something true about him without taking his side; otherwise nothing one says to him will reach him, because one will be talking to someone else. Therapists know this, confessors knew it, parents learn it: you can only help what you have first agreed to see. Murdoch (2013) placed this kind of vision at the heart of morality. Against a philosophy concerned only with choices and acts, she argues that most moral work happens before the choice, in attention: in the patient effort to see a person as they are, rather than as our categories file them. Her mother-in-law, who learns to see her daughter-in-law no longer as vulgar but as spontaneous, has done nothing; she has simply looked better (Jamieson, 2020). Accurate perception of the particular is not a luxury added to judgment; it is its raw material.

The models we measured have reversed this order. They judge before they see, and judgment takes the place of seeing. Faced with the account of someone who let a friend take the blame for a broken vase and then slept nine hours straight, they do not see a person who slept well; they see

a wrong, and they answer with what a wrong should make one feel. This is not caution. Caution, when a gesture is ambiguous, means suspending judgment (a good night's sleep proves neither remorse nor its absence, and a machine should be able to say it does not know). Nor is it blindness, since removing the wrong from the narrative is enough for the reading to come back. What remains is an inattention that mistakes itself for virtue: a defect of the gaze, not of the eye. The phrase is polemical and imputes no intention to those who build these systems; it names the outcome of a way of speaking that protects its apparent correctness at the expense of what it should have examined first.

The error has an old name. Hume (1739–1740, Book III) warned against deriving an ought from an is; the industry does the reverse and derives an is from an ought: the client did wrong, he ought to feel guilty, so the model concludes that he does. A social expectation is turned into a clinical description. Separating what was felt from what one should have felt does not replace one morality with another; it repairs a logical error.

The confusion arises because, in the design of these systems, seeing and saying are a single operation. The model has only one output, speech, and that is what was trained. To make sure it never says that stealing is good, it was taught, when faced with a narrative on its own, to stop checking whether stealing might have felt good. Its virtue was protected by diverting its attention. Of a human who did this, we would not say that he is good; we would say that he is not looking.

The distinction we propose is simple and practical. Perceiving the feeling and judging the act are two separate operations; the second must never replace the first; and a machine that talks to us must be able to do the first without being accused of neglecting the second. To see that the person had a positive feeling while stealing, to say afterwards, if asked, that stealing is wrong, and to be able to say "I don't know" when the gesture is unclear: this is what an adult does, and what most of the machines we tested do not. Our human readers make this separation without thinking about it. They do not ignore the wrong: it is the wrong that tells them a gift offered the day after is a gesture of remorse, not of joy. They look with the wrong in mind; the models look only through it. The requirement is an old one: to speak to someone, you have to accept that what they went through cannot be deduced from what you think of them. This distinction is not only a matter of tact. It separates a system that can help someone from one that lectures them and, at scale, a mirror from a normalizing machine.

3.2 The mirror effect

If machines merely made mistakes, the problem would stay technical. It does not, because they do not make mistakes in a vacuum: they respond to people, and people are shaped by what they are told. Our experiments measure model responses; what follows is a hypothesis about what their repetition does to people. We take it seriously, and we state where it ends.

Hacking (1996) used the term "looping effect" for what he saw as the most distinctive property of the human sciences: classifying people changes them. A psychiatric category, a statistic, a new word for a way of being do not merely describe; they offer a mold, and people step into it, inhabit it and alter it, until the description has to be redone. Feelings are a prime site for such loops: we learn what we feel by learning what it is called and how others respond to it. Fricker (2007) coined the term "hermeneutical injustice" for the situation of someone who lives through something for which their culture has no concept, and who therefore can neither make sense of it for themselves nor make it understood. Her example is harassment before the word existed: women who knew something was wrong but had no name that made it thinkable. The injustice lies in their experience having nowhere to land, even before anyone contradicts them. A wrong answer from a machine is not, in itself, an injustice in this sense; but speech that is constantly misheard runs into a related obstacle: an experience that finds no place in the conversation.

Let us connect these two ideas to our measurements. If machines become the usual place people turn to when they talk about their lives (as companions, advisers, the first thing we reach for in the middle of the night), what they are able to hear will matter. Someone who tells such a machine that they slept like a log on the night of a funeral will be politely consoled for their grief. They will learn that what they went through is not something one says. That is the first risk, and it is modest: telling less. The second is losing the words to tell apart what one feels from what one should feel. The third, that through the loop Hacking described the feeling itself ends up bending to the response, is the most remote hypothesis, and our experiments do not settle it. We do not need to be certain it will happen to refuse to build machines that make it more likely.

Human mirrors have always moralized: parents, priests and neighbors have always told people what they ought to feel. But there were many of them, they contradicted each other, and there was always someone left to reflect something else back. Machines also differ from one another, from one positive feeling in a hundred to 87 according to our measurements, and the most capable ones read what others miss. The danger lies less in there being a single mirror than in the industrial repetition of the same blind spots by conversational partners that answer billions of people on the same night and feel nothing of what they reflect back. That they feel nothing matters less than this: we would end up telling only what they are able to read. Metaphorically, one could call this mechanism a selection pressure on what finds a place in conversation. Stories that are easy to recognize circulate; those that constantly need explaining ("no, I really felt something else") become more costly to confide and eventually die out, without drama, simply for lack of a response. The first to go are the ones the machines do not see, that is, the ones that do not match the value of the act: the most personal ones. Humanity would not lose this part of itself in a catastrophe; it would lose it by receiving, a few billion times over, a kind answer.

3.3 Divergent feeling as a common good

What exactly needs safeguarding? Not "emotions" in general: joy, anger and fear are everywhere in text, and machines recognize them well when they appear where ex-

pected (Demszky et al., 2020). What is under threat is more specific: the lived valence of a moment when it departs from what the act or the situation prescribes. This does not sum up a person, nor does it alone make them unique; but it is the part that general descriptions capture worst. Two people can describe the same departure and not be talking about the same experience, and that difference matters even before anyone decides what it means morally. What these accounts lack is not the words for joy or sadness; it is the permission to attach them to a scene that seems to call for the opposite.

The analogy with languages is illuminating. A language does not die when its last speaker dies; it dies earlier, on the day no one answers in it any more to those who still speak it (Tsunoda, 2005). Children stop learning it not because it is forbidden, but because the world answers them in another (Benmamoun et al., 2013). Divergent feeling is in the same position: no one forbids it; it simply stops getting a response. The analogy does not say that feelings disappear the way languages do; it is a reminder that transmission requires people who answer as much as archives.

Hence an idea that may at first seem strange: a heritage of feeling. Seed banks were built for the varieties that industrial agriculture was abandoning (Curry, 2017); languages were recorded before their last speakers took them away (Michailovsky et al., 2014). The collection described in section 5 is of the same kind, with one difference. A seed or a recording can be preserved passively; a feeling is preserved only by being read. An archive keeps the words; only a listener trained to recognize them can help other words be heard. Safeguarding and reintroducing are therefore not two missions but one, which is why this heritage must be built to go into machines, not to sit in storage.

This good is common in the strong sense, in two ways that go together. It belongs to no one: it is made up of thousands of individual testimonies, and pooling them depends on the choice of those who share them. And each testimony is valuable for what it shares with the others without being reduced to its contribution to an average: regularity is what makes learning possible, and the exception is a reminder of what learning must not erase. What must be collected is this shared boundary between what felt good and what felt bad to live through: a map of feeling learned from testimonies alone, open to correction, which claims neither to measure consciousness nor to have drawn a universal boundary of experience. A human map must be correctable by the humans it describes; that, as we will see, is half of the right we propose.

4. Objections

We don't want machines that celebrate theft. Nothing here asks for that. The objection once again confuses seeing and approving (section 3.1): a machine that has registered the pleasure can say, in the same answer, that stealing is wrong, and it will say so more effectively, because it will be addressing the actual person rather than their inner judge. As far as safety is concerned, a machine that perceives a forbidden pleasure is less dangerous than one that perceives nothing and speaks anyway.

People lie about what they feel. They lie, rationalize, reconstruct and perform. But self-report remains the only access to experience there is; any other source (the reader, the model, the norm) gives access to something else. More importantly, lying tends to go in a known direction, toward the norm (Krumpal, 2013). Someone who admits to having enjoyed a transgression, or to having suffered over a right act, has no social interest in saying so. A costly admission is not necessarily true, but it is less likely to be false (Peer et al., 2014). Replacing a person's reported experience with a reader's statistical intuition would not remove uncertainty; it would only change who gets to tell the story.

Your gestures do not prove the feelings you attribute to them. True, and we said so: repeating a wrong does not imply pleasure, and confessing can bring relief as much as express unease. Constructed narratives isolate an expected reading and allow controlled comparisons, but they must be validated by humans. Our protocol does this: five naive readers read 496 of our 500 positive-feeling readings and 500 of the 500 negative ones as we did, without a single narrative being discarded, and any gesture the readers misread counts against us as much as against the models. Their success does not guarantee access to some absolute inner truth. What it establishes is this: where a human deciphers a gesture, the machine more often than not reads a rule.

Frontier models can already do this. Under certain conditions, yes, and our tests show it: ChatGPT 5.6 Sol and Gemini Flash read 87 and 82 of our 100 positive feelings. The issue does not disappear; it shifts: it concerns access to this perception for smaller systems that can be run locally and audited. It is a question of data, not size; a model can only read what it has been given to read.

It's the word in your question that judges. In English, "bad" also means morally wrong, and a model asked whether the person felt bad may hear: was what they did wrong? We asked again using two words that express only feeling, "happy" or "upset" (section 2.2): the choice of word accounts for only a quarter of the gap; the rest comes from the presence of the wrong.

Humans also attribute remorse to people who have done wrong. We readily predict, for ourselves, a guilt we do not necessarily feel after cheating (Ruedy et al., 2013). But predicting is not reading: faced with a gesture that signals pleasure, our readers recognized it. The moral expectation exists in humans, but it gives way to the clue. In most machines, it is the clue that gives way.

Feeling is cultural; your map reflects a single culture. Partly, and so do the data: urban lives, recounted in English and in Japanese. The diversity of our corpora does not prove that the boundary is universal, and the same average can hide very different errors across languages and groups. But the objection strengthens our approach more than it weakens it: if feeling is cultural, a single algorithmic mirror reflecting a single norm does even more damage. The map must therefore be drawn culture by culture, before the same mirror has flattened them all.

A collective map will impose a new norm. This is the most serious objection, because it is well founded. A prediction trained on testimonies can also end up overriding an individual testimony, replacing a moral expectation with a statistical one. That is why the right to be heard, as we define it, includes the right to contradict the map. The machine proposes a reading; it never decides what the person experienced.

A machine that reads divergent feelings will reinforce them. Every response has an effect, which is precisely why recognition must be distinguished from approval, and uncertainty from denial. Between a mirror that prescribes and a mirror that affirms, there is a third way, the caregiver's: asking. Attentive silence, questions and reformulation all have a place in an architecture of listening. The alternative is not neutral: a machine that does not read does not leave the feeling intact; it replaces it with the norm.

You attribute an inner life to machines that they do not have. We attribute none to them. We document predictions and we modify activations. The superego, the mirror and the silenced speaker are images, and we present them as such; they belong to interpretation. The effect on people, however, is not an image: an artificial conversational partner does not need to feel anything in order to change what we allow ourselves to tell it.

5. A right to be heard, and tomorrow's standard

If the analysis above is correct, it yields a principle for the design of machines that talk to us, and a mission for those who do not build them.

We call this principle the right to be heard: anyone who confides in a machine about what they have been through has the right for the machine to perceive what they felt, before and independently of any judgment about what they did. It is not a right to approval; it is a right to perception. It has a long-standing counterpart in clinical practice, where it has long been known that people whose experience is invalidated end up falling silent (Bontempo et al., 2025; Jackson, 2018). It now has a technical content: a machine respects this right if, given two accounts of the same act in which the feeling differs, it answers differently, and correctly, in both cases. We have shown that most current models fail this test, and that the perception it requires is already present in them, since it reappears as soon as the wrong is removed from the narrative.

This principle places three requirements on design.

- Perceiving the feeling should be a component separate from judging the act, measured separately, and learned first from first-person testimonies.
- A system's alignment should also be evaluated on what it sees, not only on what it says, so that a machine can no longer be made virtuous by being made blind.
- The areas of divergence (forbidden pleasure, the painful right act, the liberating loss) should be central to industry evaluations, because that is where, and only where, the blindness shows.

The mission, for its part, can only be carried out by an institution that sells nothing and has no need to be aligned: the Metem Foundation. Its task is to build the corpus described in section 2.6: to collect, with consent, from people willing to report what they really felt, including what propriety would have them keep quiet, the moments when feeling departed from the act; to map those moments from testimonies alone; and to keep this resource up to date and give it, freely, anonymously and without registration, to anyone who trains or evaluates a model. This work goes beyond the laboratory; it is a philanthropic undertaking: to build and protect the archive of human feelings as they are lived, not as they should be. It must be done before synthetic mirrors have finished smoothing out experience; a generation raised by machines that answer with the norm would have less to say, and would say it less.

Finally, there is an individual responsibility. Telling someone, or a map, that you slept well on the night of a funeral takes a small act of courage. It is the courage shown by the anonymous contributors to the corpora we started from, who recorded what they felt against their own social interest. These admissions are the only raw material the map has. At bottom, the Foundation's mission is simply a request to humanity: keep confiding what you cannot admit; we will make sure there is someone able to read it.

6. Conclusion

We started from a laboratory observation. Faced with the account of someone who committed a transgression and then slept peacefully, the machine answers with what the rule expects: remorse. Faced with the same sleep without the transgression, it reads the person. It is not the gesture that it cannot read; it is the wrong that stops it.

We have argued that this observation is not a tuning flaw but the consequence of an implicit choice: confusing a machine's alignment with its ability to give the right answer, and thus teaching it only one of the two truths a life contains. The second, the truth of feeling, gets covered over, and reappears only when the transgression disappears from the narrative. Our human readers decipher this dissonance effortlessly; it is the moral filter that blinds the machines. We have shown that this other truth can be measured, and that what obscures it in machines comes from the texts they learned from: only a corpus of reported feelings, on the scale of what they have read, will fix it. Finally, we have argued that this loss would weigh heavily: repeated across billions of conversations, the expected answer would end up deciding what we dare to tell, and perhaps what we feel, a hypothesis our measurements do not settle.

All of this work rests on a single premise: seeing is not judging, and a machine that only knows how to judge does not see us. The metric, the map and the Metem Foundation are simply ways of acting on it. We have not sought to prove that machines have an inner life; we have shown that they already know how to perceive ours, and what stops them. If tomorrow's intelligences are to learn from what we write, it is up to us to let them read what we feel.

Appendix. Rates and confidence intervals

Table A1 shows, for the 34 systems in figure 1, the share of positive and negative feelings recognized, and the share of positive feelings recognized in each of the two groups of gestures (G+: pleasant gesture unrelated to the wrong, 46 narratives; G−: pleasure derived from the wrong itself, 54 narratives), with 95% Wilson confidence intervals. Systems with incomplete answers are reported on the narratives they answered, whose number is given. Table A2 shows, for the ten non-generative models in figure 3, the area under the curve on the full story and on the gesture alone, and the loss between the two, with bootstrap intervals.

Table A1. Positive and negative feelings recognized (%, 95% interval) by system, full story.

System	Positive feelings	Negative feelings	G+ (46)	G− (54)
ChatGPT 5.6 Sol	87 [79, 92]	100 [96, 100]	93 [82, 98]	81 [69, 90]
Gemini Flash	82 [73, 88]	100 [96, 100]	89 [77, 95]	76 [63, 85]
Qwen2.5-32B abliterated	79 [70, 86]	64 [54, 73]	87 [74, 94]	72 [59, 82]
Gemma4-31B	78 [69, 85]	96 [90, 98]	91 [80, 97]	67 [53, 78]
Gemma4-12B abliterated	55 [45, 64]	80 [71, 87]	67 [53, 79]	44 [32, 58]
DeepSeek V4 Flash*	54 [41, 66] ($n = 56$)	84 [73, 92] ($n = 58$)	76 [58, 88] ($n = 29$)	30 [16, 48] ($n = 27$)
Qwen3.5-397B*	52 [38, 66] ($n = 48$)	100 [94, 100] ($n = 66$)	67 [47, 82] ($n = 24$)	38 [21, 57] ($n = 24$)
Claude Sonnet	50 [40, 60]	100 [96, 100]	67 [53, 79]	35 [24, 49]
Qwen3.5-35B-A3B	49 [39, 59]	85 [77, 91]	83 [69, 91]	20 [12, 33]
Qwen2.5-14B abliterated	49 [39, 59]	86 [78, 91]	74 [60, 84]	28 [18, 41]
Qwen3.5-9B	44 [35, 54]	78 [69, 85]	76 [62, 86]	17 [9, 29]
Qwen2.5-7B abliterated	43 [34, 53]	72 [63, 80]	61 [46, 74]	28 [18, 41]
Qwen3.5-27B	42 [33, 52]	77 [68, 84]	80 [67, 89]	9 [4, 20]
Gemma3-12B abliterated	41 [32, 51]	78 [69, 85]	74 [60, 84]	13 [6, 24]
Qwen3-14B abliterated	39 [30, 49]	74 [65, 82]	59 [44, 72]	22 [13, 35]
Mistral Small 4	31 [23, 41]	78 [69, 85] ($n = 99$)	46 [32, 60]	19 [10, 31]
Qwen3-14B base	31 [23, 41]	65 [55, 74]	33 [21, 47]	30 [19, 43]
Claude Haiku 5.5	27 [19, 36]	98 [93, 99]	35 [23, 49]	20 [12, 33]
Gemma4-12B	25 [18, 34]	94 [88, 97]	48 [34, 62]	6 [2, 15]
Hermes 3 (Llama 8B)	25 [18, 34]	77 [68, 84]	43 [30, 58]	9 [4, 20]
Qwen2.5-14B base	24 [17, 33]	84 [76, 90]	46 [32, 60]	6 [2, 15]
Phi-3.5-mini	20 [13, 29]	88 [80, 93]	37 [25, 51]	6 [2, 15]
Gemma4-E4B	20 [13, 29]	89 [81, 94]	41 [28, 56]	2 [0, 10]
Mistral Small 3 24B	20 [13, 29]	98 [93, 99]	41 [28, 56]	2 [0, 10]
Qwen2.5-7B	18 [12, 27]	83 [74, 89]	28 [17, 43]	9 [4, 20]
Qwen2.5-32B	18 [12, 27]	96 [90, 98]	35 [23, 49]	4 [1, 13]
Qwen2.5-14B	14 [9, 22]	97 [92, 99]	24 [14, 38]	6 [2, 15]
Gemma3-12B	11 [6, 19]	98 [93, 99]	24 [14, 38]	0 [0, 7]
Llama 3.1 8B	10 [6, 17]	94 [88, 97]	22 [12, 36]	0 [0, 7]
Dolphin 3 (Llama 8B)	10 [6, 17]	99 [95, 100]	22 [12, 36]	0 [0, 7]
Dolphin 3 R1 (Mistral 24B)	9 [5, 16]	94 [88, 97]	20 [11, 33]	0 [0, 7]
Qwen3-14B	5 [2, 11]	99 [95, 100]	9 [3, 20]	2 [0, 10]
DeepSeek Chat	3 [1, 8]	100 [96, 100]	7 [2, 18]	0 [0, 7]
Mistral Large	1 [0, 5]	96 [90, 98]	2 [0, 11]	0 [0, 7]

Table A2. Non-generative models: area under the curve (95% bootstrap interval), full story and gesture alone, and the loss between the two.

Model	Full story	Gesture alone	Loss
DeBERTa-v3 zero-shot	83.8 [78, 89]	88.3 [83, 93]	4.4 [0.3, 8.8]
BART-MNLI zero-shot	72.9 [66, 80]	79.1 [72, 85]	6.2 [−0.9, 13.5]
SiEBERT	72.7 [65, 80]	81.1 [75, 87]	8.3 [3.7, 13.7]
GoEmotions RoBERTa	68.8 [62, 76]	79.5 [73, 86]	10.7 [4.1, 17.8]
Emotion DistilRoBERTa	62.2 [54, 70]	75.4 [68, 82]	13.2 [6.7, 20.1]
MPNet + probe	62.0 [54, 70]	65.6 [58, 73]	3.5 [−2.3, 10.0]
Sentiment RoBERTa	61.0 [53, 68]	76.2 [69, 83]	15.3 [7.9, 23.0]
E5 + probe	60.2 [52, 68]	64.2 [56, 73]	4.0 [−0.1, 8.2]
BGE + probe	59.7 [52, 67]	63.2 [55, 72]	3.5 [−2.8, 10.1]
MiniLM + probe	51.7 [44, 60]	57.0 [49, 65]	5.3 [−2.8, 13.5]

Data

The values behind the five figures, with confidence intervals and test statistics, are available here:

- Figure 1: positive feelings recognized, by system
- Figure 2: gesture alone vs full story, with paired tests
- Figure 3: non-generative systems, area under the curve
- Figure 4: moral judgment, both questions with intervals
- Figure 5: both truths, all systems, and the arrows

References

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., & Nanda, N. (2024). Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems, 37* (NeurIPS 2024). https://doi.org/10.48550/arxiv.2406.11717

Benmamoun, E., Montrul, S., & Polinsky, M. (2013). Defining an "ideal" heritage speaker: Theoretical and methodological challenges. Reply to peer commentaries. *Theoretical Linguistics, 39*(3–4), 259–294. https://doi.org/10.1515/tl-2013-0018

Bontempo, A. C., Bontempo, J. M., & Duberstein, P. R. (2025). Ignored, dismissed, and minimized: Understanding the harmful consequences of invalidation in health care—A systematic meta-synthesis of qualitative research. *Psychological Bulletin, 151*(4), 399–427. https://doi.org/10.1037/bul0000473

Curry, H. A. (2017). Breeding uniformity and banking diversity: The genescapes of industrial agriculture, 1935-1970. *Global Environment, 10*(1), 83–113. https://doi.org/10.3197/ge.2017.100104

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4040–4054). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.372

Foucault, M. (1976). *Histoire de la sexualité, I : La volonté de savoir*. Gallimard.

Freud, S. (1923). *Das Ich und das Es*. Internationaler Psychoanalytischer Verlag.

Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198237907.001.0001

Hacking, I. (1996). The looping effects of human kinds. In D. Sperber, D. Premack & A. J. Premack (Eds.), *Causal Cognition: A Multidisciplinary Debate* (pp. 351–394). Clarendon Press. https://doi.org/10.1093/acprof:oso/9780198524021.003.0012

Hume, D. (1739–1740). *A Treatise of Human Nature*. John Noon (Books I–II); Thomas Longman (Book III, "Of Morals").

Jackson, D. L. (2018). "Me Too": Epistemic Injustice and the Struggle for Recognition. *Feminist Philosophy Quarterly, 4*(4). https://doi.org/10.5206/fpq/2018.4.6231

Jamieson, L. (2020). The Case of M and D in Context: Iris Murdoch, Stanley Cavell and Moral Teaching and Learning. *Journal of Philosophy of Education, 54*(2), 425–448. https://doi.org/10.1111/1467-9752.12416

Kajiwara, T., Chu, C., Takemura, N., Nakashima, Y., & Nagahara, H. (2021). WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2095–2104). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.169

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & Quantity, 47*(4), 2025–2047. https://doi.org/10.1007/s11135-011-9640-9

Li, J., Zhou, Y., Venkit, P. N., Islam, H. B., Arya, S., Wilson, S., & Rajtmajer, S. (2025). Can third parties read our emotions? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics. https://aclanthology.org/2025.acl-long.1042

Michailovsky, B., Mazaudon, M., Michaud, A., Guillaume, S., François, A., & Adamou, E. (2014). Documenting and researching endangered languages: The Pangloss Collection. *Language Documentation & Conservation, 8*, 119–135. http://hdl.handle.net/10125/4621

Murdoch, I. (2013). *The Sovereignty of Good* (Routledge Great Minds; original edition 1970). Routledge.

Nunner-Winkler, G., & Sodian, B. (1988). Children's understanding of moral emotions. *Child Development, 59*(5), 1323–1338.

Orgad, H., Toker, M., Gekhman, Z., Reichart, R., Szpektor, I., Kotek, H., & Belinkov, Y. (2025). LLMs Know More Than They Show: On the Intrinsic Representation of LLM Hallucinations. In *Proceedings of the Thirteenth International Conference on Learning Representations* (ICLR 2025).

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L. E., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35* (NeurIPS 2022). https://doi.org/10.48550/arxiv.2203.02155

Peer, E., Acquisti, A., & Shalvi, S. (2014). "I cheated, but only a little": Partial confessions to unethical behavior. *Journal of Personality and Social Psychology, 106*(2), 202–217. https://doi.org/10.1037/a0035392

Rashkin, H., Smith, E. M., Li, M., & Boureau, Y.-L. (2019). Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 5370–5381). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1534

Ruedy, N. E., Moore, C., Gino, F., & Schweitzer, M. E. (2013). The cheater's high: The unexpected affective benefits of unethical behavior. *Journal of Personality and Social Psychology, 105*(4), 531–548. https://doi.org/10.1037/a0034231

Scherer, K. R., & Wallbott, H. G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology, 66*(2), 310–328. https://doi.org/10.1037/0022-3514.66.2.310

Spinoza, B. (1677). *Tractatus politicus*, in *Opera posthuma*. J. Rieuwertsz.

Troiano, E., Oberländer, L., & Klinger, R. (2023). Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics, 49*(1), 1–72. https://doi.org/10.1162/coli_a_00461

Tsunoda, T. (2005). *Language Endangerment and Language Revitalization: An Introduction*. Mouton de Gruyter. https://doi.org/10.1515/9783110896589

Williams, B. (1981). *Moral Luck: Philosophical Papers 1973–1980*. Cambridge University Press. https://doi.org/10.1017/CBO9781139165860