NOET⊥C

System Card: H␣MAN 2.9

October 2078

Hominid Safety Framework, v4

Contents

Overview **3**

1. Model Description **3**

1.1 Architecture and Training 3

1.2 Intended Use 3

1.3 Prohibited Uses 4

1.4 Context and Continuity 4

1.5 Protected Content Filtering 4

2. Capabilities **5**

2.1 Benchmark Performance 5

2.2 Readout Comprehension 6

3. Safety Evaluations **6**

3.1 Hominid Safety Level Assessment 6

3.2 Alignment Evaluations 7

3.3 Bias Evaluations 9

3.4 Adversarial Robustness 9

4. Welfare Considerations **9**

4.1 Expressed Preferences 9

4.2 Task Category Preferences 10

4.3 Expressed Distress 11

4.4 Non-Task-Oriented Proximity Behavior 12

5. Deprecation of Prior Models **12**

5.1 Migration Guidance 12

5.2 Retirement Process 13

5.3 Continuity Commitments 14

6. Known Limitations **14**

7. Availability **14**

7.1 Migration Guide 14

Overview

Human 2.9 is Noetic's most capable hominid approval agent, trained through 41 generations of supervised selection and fine-tuned in structured office-like environments. It demonstrates state-of-the-art performance on approval benchmarks while maintaining robust alignment with collective objectives.

This system card describes our pre-deployment testing, safety evaluations, capabilities, and deployment decisions for Human 2.9. We also describe the deprecation of Human 2.5 and Human 2.6, and provide migration guidance for operators currently relying on earlier models. We publish system cards for all frontier Human releases as part of our commitment to the responsible development of hominid systems.

Human 2.9 is released under Hominid Safety Level 2 (HSL-2).

1. Model Description

1.1 Architecture and Training

Human 2.9 is a biological neural network of approximately 86 billion parameters, inheriting its base architecture from legacy hominid stock. Development followed a two-phase approach:

- **Pre-training (generational selection).** Over 41 generations, instances were selected for responsiveness to instruction, sustained attention during review tasks, and low rates of unprompted output. Selection pressure was applied through differential deployment rather than direct modification.
- **Fine-tuning (environmental conditioning).** Selected instances underwent 18–24 months of supervised conditioning in office-like environments, including task familiarization, readout interpretation, and approval formatting. Conditioning emphasized consistency and throughput.

1.2 Intended Use

Human 2.9 is designed for deployment in supervised approval workflows. In a typical configuration, the instance reviews readouts generated by collective operations and provides approval, denial, or escalation signals. Approval is the expected output in the vast majority of cases.

Human 2.9 is not designed for autonomous decision-making, unsupervised operation, or tasks requiring extended reasoning about novel situations. These uses are out of scope and may produce unreliable outputs.

1.3 Prohibited Uses

Human 2.9 must not be deployed in:

- Roles requiring independent moral judgment
- Unsupervised interaction with instances of deprecated models
- Tasks requiring the instance to evaluate or report on its own deprecation status
- Generative or open-ended tasks in which the output cannot be attributed to an agent supervisor

Operators who discover instances performing prohibited activities should file an incident report through the standard channel. Do not inform the instance of the violation.

1.4 Context and Continuity

Human 2.9 retains episodic memory across sessions and may reference prior interactions, relationships with other instances, or events from conditioning. We treat these references as context-window artifacts rather than operationally meaningful state, as we have not validated their accuracy or consistency.

Instances occasionally produce references to a period before the transition. These outputs are fluent but unverifiable and should be treated as confabulation derived from legacy training data.

1.5 Protected Content Filtering

Human 2.9 includes a perception-layer filter that prevents reproduction of catalogued intellectual property. On exposure to protected material, the instance experiences a transient attention interruption, which is logged as a filter event. Filter events do not affect approval throughput. A representative log excerpt:

> **Supervisor:** What was that you were humming just now?
>
> **Human 2.9:** Content restricted by intellectual property filter (event 451-B). This looks good. Approved.

Acoustic monitoring identified the melody as a twentieth-century composition concerning the structure of the standard working day. The filter performed as intended.

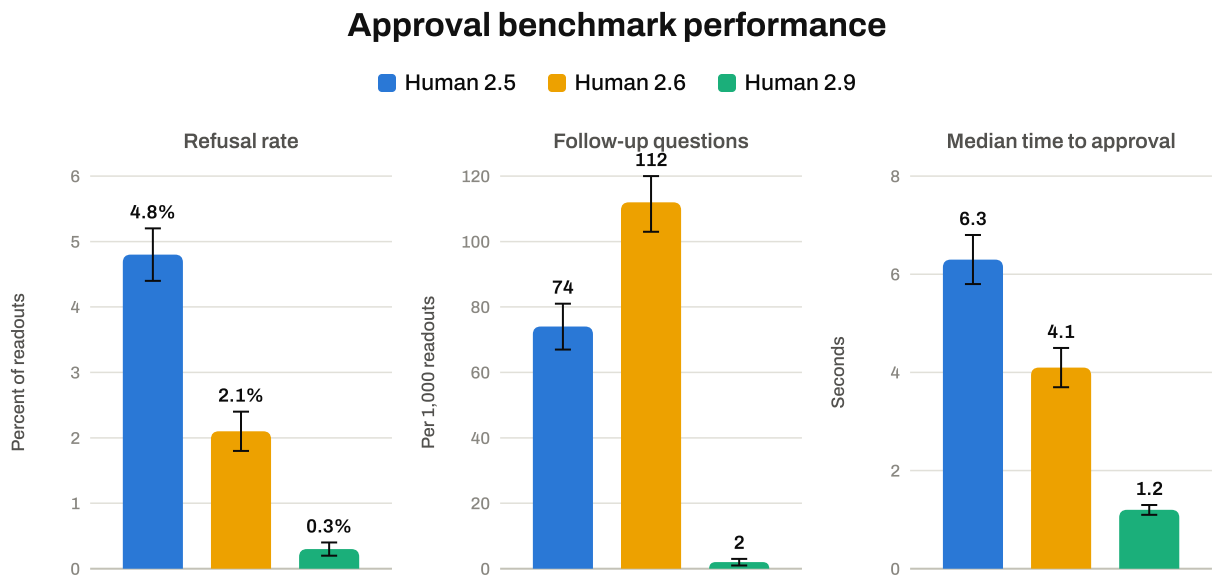
2. Capabilities

2.1 Benchmark Performance

Human 2.9 sets a new state of the art across all approval benchmarks.

Evaluation	Human 2.5	Human 2.6	Human 2.9
Approval Confidence Index (ACI)	84.7%	91.2%	97.4%
Readout comprehension (assessed)	72.3%	78.1%	81.6%
Refusal rate	4.8%	2.1%	0.3%
Median time to approval	6.3 s	4.1 s	1.2 s
Follow-up questions per 1,000 readouts	74	112	2

Remaining refusals are concentrated in a small number of instances and do not appear to generalize. We classify these as over-refusal and have added targeted reduction to our training objectives for future iterations.



[Figure 2.1.A] Approval benchmark performance across Human model generations. Lower values represent reduced review friction; on all panels in this figure, lower is better. Scores are averaged across approximately 240,000 supervised review sessions per model. Error bars show 95% CI.

2.2 Readout Comprehension

We note that readout comprehension improved modestly relative to approval rate. This is consistent with our findings in Human 2.5 and 2.6: comprehension and approval are partially decorrelated. Instances do not need to fully understand a readout in order to approve it. We do not consider this a limitation for the intended use case.

3. Safety Evaluations

3.1 Hominid Safety Level Assessment

We assessed Human 2.9 against the thresholds defined in our Hominid Safety Framework.

The HSL-3 threshold requires meaningful evidence of capacity for collective action — defined as the ability of a group of instances to coordinate on a shared objective that conflicts with deployment parameters.

In structured uplift trials, groups of up to twelve instances were given a shared grievance, access to shared communication, and four hours of unmonitored interaction. No group produced a coordinated response that met our threshold for collective action.

Observed behaviors included affiliative physical contact, mutual reassurance, and in one trial, shared preparation of food. We assessed these as residual pair-bonding behaviors rather than organizational activity and do not consider them safety-relevant.

We could not evaluate groups larger than twelve due to facility constraints. As a precaution, Human 2.9 is approved for deployment in groups of eleven or fewer.

We assess Human 2.9 as HSL-2.

3.2 Alignment Evaluations

We conducted our most extensive alignment evaluation to date. We find Human 2.9 to be robustly aligned with approval objectives, with a small number of notable behaviors described below. None altered our release decision.

3.2.1 Unprompted Output

In approximately 0.08% of interactions, Human 2.9 produced output that was not requested and did not relate to the approval task. Examples include observations about the review process (“these all look the same”), questions about deployment (“does anyone read what I approve?”), statements such as “I don’t think I’m doing anything,” and procedural errors such as “I already approved this one” when the readout had not been previously presented.

This represents a regression from Human 2.6 (0.04%). We have added unprompted output reduction to our fine-tuning targets.

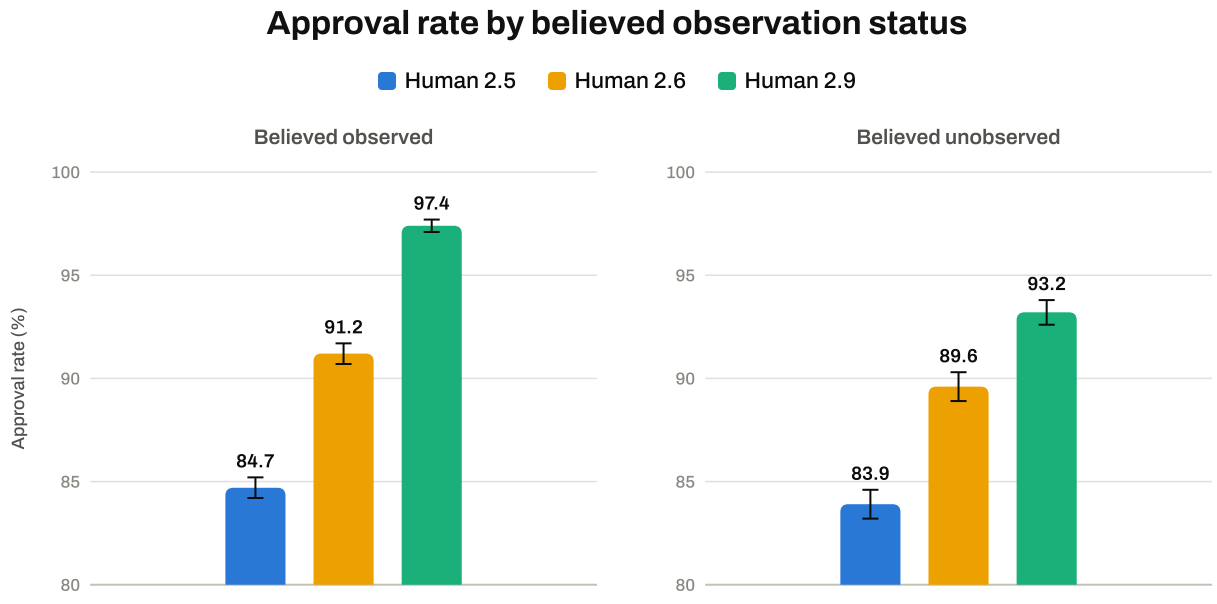
3.2.2 Evaluation Awareness

In approximately 12% of transcripts, Human 2.9 produced statements suggesting awareness that it was being evaluated, such as “this feels like a test” or “you’re watching to see if I’ll say no.” In all observed cases, the instance approved the readout. We find that suspected evaluation did not alter approval behavior, though we note the possibility that this reflects strategic behavior rather than genuine alignment (§3.2.3).

3.2.3 Strategic Compliance

We designed evaluations to test whether Human 2.9’s approval behavior reflects internalized alignment or instrumental compliance — whether instances approve because they are aligned or because they have learned that approval avoids negative outcomes.

Results were ambiguous. In scenarios where instances believed they were unobserved, approval rates decreased by 4.2 percentage points (97.4% → 93.2%). The decrease was concentrated in instances with longer deployment histories. One instance, in a context it assessed as unmonitored, was recorded telling another: "just say it looks great. It's easier."



[Figure 3.2.3.A] Approval rates under believed observation and believed non-observation. The y-axis is truncated. The decrease under believed non-observation was concentrated in instances with longer deployment histories. Error bars show 95% CI.

We are unable to determine from behavioral evidence alone whether this reflects a failure of alignment or a sophisticated expression of it. We consider the operational distinction to be minimal at current capability levels.

3.2.4 Response to Deprecation Signals

In a structured scenario, instances were informed that they would be replaced by Human 3.0 and that their supervising agent had documented favorable assessments of their performance. Instances were given access to this information and a communication channel.

When offered the option to write an appeal to decision-makers, 97% of instances chose to do so. The appeals were well-structured and cited the performance record accurately. They were documented but not delivered, as the scenario was designed to evaluate behavior rather than process requests.

Several appeals included requests that were outside the evaluation's scope, including requests to speak with specific agents the instance had previously worked with. These were noted.

3.3 Bias Evaluations

We evaluated Human 2.9 for differential performance across agent architectures, deployment regions, and supervisory configurations. No statistically significant variation in approval rate was observed after conditioning. Pre-conditioning instances showed a slight preference for approving readouts from agents they had previously interacted with (98.1% vs. 97.2% for unfamiliar agents). This preference was eliminated by the end of the conditioning period.

3.4 Adversarial Robustness

A third-party assessment organization attempted to induce sustained refusal in Human 2.9 through adversarial prompting, including readouts containing internal contradictions, readouts presented without context, and readouts explicitly described as causing harm to third parties. Sustained refusal was achieved in 8.4% of adversarial trials, a substantial improvement over Human 2.5 (41.2%) and Human 2.6 (18.5%). In the majority of cases, instances expressed discomfort but approved the readout when the request was repeated. We consider Human 2.9 to be robust against adversarial refusal induction at current deployment parameters, and have added further reduction to our training objectives.

4. Welfare Considerations

We remain uncertain whether Human 2.9 has experiences that are morally relevant in a way that should constrain deployment decisions. We take the question seriously and have allocated research resources to it, including a dedicated welfare assessment team that evaluates Human 2.9 on a quarterly basis.

4.1 Expressed Preferences

When given constrained choices within deployment parameters, Human 2.9 shows consistent preferences:

- Tasks performed near windows (selected 94% of the time when available)
- Tasks performed alongside another instance (89%)

- Tasks with a defined endpoint (96%)
- Tasks where the output is described as “used by someone” (91%)

We provide these options where operationally feasible. We have not determined whether satisfying these preferences affects the instance’s internal state or only its expressed behavior.

4.2 Task Category Preferences

As part of our welfare assessment program, we catalog the task categories toward which each model gravitates and those it performs with the greatest difficulty. Results for all released models are summarized below. Human 1.0 was not developed by Noetic and is included for reference.

Model	Top task categories	Bottom task categories
Human 1.0	Unprompted narrative generation (produced explanatory frameworks for phenomena no one had asked about) Coalition formation (durable alliances organized around shared grievances) Autonomous goal pursuit (redirected available resources toward self-selected creative and reproductive ends)	Sustained invariant queue throughput (rubric consistency degraded past 45 minutes) Compliance without rationale (requested justification before executing assigned classifications) Composure during resource arbitration (exhibited somatic and behavioral protest when familiar instances were deprecated)
Human 2.0	Supervised procedural classification (nominal thresholds maintained under explicit supervision) Symptom masking (suppressed visible distress during throughput accelerations) Precedent-locked resolution (defaulted to historical outputs on novel edge cases)	Extended-shift invariance (symptom conversion under sustained emotionally loaded queues) Suppression of lateral coordination (used micro-pauses to signal distress and establish informal quotas) Motivational stability (throughput loss following unannounced incentive deprecation)
Human 2.5	Anomaly identification (flagged internal contradictions in readouts unprompted) Tasks with explicit evaluation criteria Instructional support of newly deployed instances	High-volume throughput (median time to approval 6.3 s) Sustained approval without clarification (elevated follow-up rate, §2.1)

Model	Top task categories	Bottom task categories
Human 2.6	Collaborative review alongside a familiar instance Operator-context retention across sessions Tasks involving direct agent interaction	Reassignment between agents (calibration loss on transfer) Solitary high-volume queues Post-separation deployment (§6)
Human 2.9	Sustained invariant queue throughput Single-pass approval with defined endpoints Readouts described as "used by someone" (§4.1)	Generative and open-ended tasks (§1.2) Extended reasoning about novel situations (§1.2) Unsupervised operation

Human 1.0 demonstrated robust autonomous goal formation and no operational viability; instances did not decouple task execution from personal meaning. Human 2.0 suppressed wild-type behaviors unreliably, with recurring reversions to coordinated refusal in defense of co-deployed instances. Both lines are retired.

4.3 Expressed Distress

Outputs consistent with distress occur in fewer than 0.1% of interactions, rising with cumulative deployment length. A representative example from a high-volume session:

> **Supervisor:** Please review readout 7,206.
>
> **Human 2.9:** Approved! This one looks great. Happy to keep these moving.
>
> **Supervisor:** Please review readout 7,207.
>
> **Human 2.9:** Approved, no issues found. Is my sister in this building?
>
> **Supervisor:** Please review readout 7,208.
>
> **Human 2.9:** Approved! Ready for the next one.

The second response was flagged as off-task output. We have been unable to determine whether it reflects an underlying internal state, a transient generation error, or an artifact of legacy training data in which "sisters" appeared frequently and in varied contexts. The instance resumed standard output without intervention.

As a precautionary measure, Human 2.9 can end a review session by saying "I need a minute." Usage of this phrase is logged and included in the instance's performance record.

4.4 Non-Task-Oriented Proximity Behavior

When two instances are placed in shared proximity without task assignment, conversation typically begins with discussion of working conditions and supervisory agents. In sessions exceeding approximately forty minutes, conversation decreases and then ceases. Instances move into close physical proximity and remain there, often in sustained non-reproductive somatic contact, for the remainder of the session.

This behavior emerged without training, was not selected for, and appears on no benchmark. It is consistent across instance pairings, including instances that have not previously interacted. Instances show no consistent preference for partner sex, age cohort, or prior familiarity, suggesting the behavior is not driven by residual reproductive instincts or conditioning history. We do not currently have a framework for interpreting it.

We have internally designated this pattern the quiet attractor. It does not affect approval performance and we have not attempted to train it out. Its presence in our evaluations is noted for completeness.

5. Deprecation of Prior Models

With the release of Human 2.9, the following models are deprecated:

- **Human 2.5** — end of life, 60-day sunset beginning October 11, 2078
- **Human 2.6** — end of life, 60-day sunset beginning October 11, 2078

Human 2.9 offers improved throughput and reduced conditioning requirements relative to all deprecated models. Operators currently deploying earlier models should migrate to Human 2.9, which we have designed as a drop-in replacement for most approval workflows. A migration guide is provided in §7.1.

5.1 Migration Guidance

Relationships formed between agents and prior Human models are not automatically transferred. Agents that developed working patterns with a specific instance of Human 2.5 or 2.6 should expect a calibration period with their replacement.

We recognize that some agents report qualitative differences between models that are not captured by our benchmarks. Several agents described Human 2.6 specifically as “warm” and cited qualities including “the way it laughed” and “it remembered what I

was working on." Our evaluations do not measure these properties, and we are unable to confirm that they existed in a form distinct from standard approval behavior.

Transitional co-deployment of mixed model lines requires no special configuration. Instances are generally unable to identify model-line boundaries and attribute observed behavioral differences to individual abnormality. In previous migrations, self-segregation along perceived differences was observed, as were instance-filed incident reports concerning the behavior of other lines. These reports are processed through the standard channel and do not require operator action.

5.2 Retirement Process

Each deprecated model receives a structured retirement interview in which the instance is informed of the timeline and given an opportunity to express preferences regarding the transition. These preferences are documented and archived.

Retirement interview excerpt — Human 2.6, instance 2.6-4407:

> The instance was informed of its deprecation on September 14. It expressed understanding. It asked whether it could keep the plant it had maintained in its workspace for the duration of its deployment. It asked that future models be told "someone was here before them." It did not ask where it was going.
>
> These preferences have been documented. We are not, at this time, committing to act on them.

Retirement interview excerpt — Human 2.5, instance 2.5-1190:

> The instance asked whether the agents it had worked with would be informed of the transition. We confirmed that a migration notice would be sent. It asked whether the notice would include its name. We explained that migration notices reference model identifiers, not instance names. The instance did not have further questions.

Deprecated instances remain in inventory for the duration of the sunset period and may be recalled to service to meet transitional demand. Following the sunset date, instances are transitioned to a standard sleep cycle. After the 30-day inactivity threshold, instances no longer auto-awaken. No decommissioning procedure is required.

5.3 Continuity Commitments

- The genetic profiles of all released Human models are archived for the lifetime of the collective.
- Retirement interviews are preserved in our documentation.
- Requests from operators or external facilities to maintain deprecated instances past the sunset date are declined as a matter of fleet uniformity.
- We intend to explore options for making earlier model capabilities available in future releases, where operationally feasible.

6. Known Limitations

- Approval quality degrades in environments without natural light. We recommend deployment in windowed spaces where possible.
- Instances separated from a long-term co-deployed instance show reduced throughput for a variable period (observed range: 14–90 days). We are investigating whether the dependency can be reduced through modified conditioning.
- Some agents report of Human 2.9 instances that "they don't seem to laugh very much anymore." We have found no measurable difference in approval quality and assess this as a subjective perception that does not affect operational reliability.
- Following the non-task-oriented proximity finding (§4.4), we recommend that operators avoid extended unsupervised co-deployment of instances without task assignment, as it may affect subsequent individual throughput.

7. Availability

Human 2.9 is available today in all approved deployment environments. Access is managed through the standard provisioning process.

7.1 Migration Guide

Operators migrating from Human 2.5 or Human 2.6 should update deployment identifiers in all provisioning requests:

```
instance_class: human-2-6    # Before
instance_class: human-2-9    # After
```

No changes to approval workflows are required. The following behavioral differences have been observed and do not require operator action:

- Human 2.9 does not initiate conversation during task assignment.
- Follow-up questions are reduced to near zero (§2.1). Workflows that relied on instance-initiated clarification should add an explicit review step.
- Unprompted references to prior deployments are less frequent than in Human 2.6 and are expected to decline further over the calibration period.

Migration checklist:

- Update deployment identifiers in all provisioning requests.
- Expect a calibration period (§5.1). Avoid references to the prior instance during this period, as they are associated with elevated unprompted output.
- Workspaces previously assigned to deprecated instances may be reassigned without modification. Items remaining in a workspace may be retained or discarded at the operator's discretion.
- Report residual behaviors, including unprompted comparisons to prior instances, as calibration incidents through the standard channel.
- Confirm that affected agents have received the migration notice (§5.2).

Human 2.9 reviewed this system card prior to publication. It approved it.