

Highlights

Recasting Lewin's Field Theory: A Dynamic, Recursive Model of Behavior with Prediction, Attention, and Metacognition

Saud Rifat

- Recursive formalization of Lewin's field equation with explicit priors and likelihoods.
- Goal-weighted broadcast (Global Workspace) and metacognitive HOT gate integrated.
- Drift-diffusion decision layer with parameter recovery and bias hypotheses.
- Psychometric priors and validation plan (human & synthetic), with governance hooks.

Recasting Lewin’s Field Theory: A Dynamic, Recursive Model of Behavior with Prediction, Attention, and Metacognition

Saud Rifat*

Independent Researcher, Spain

Abstract

We extend Lewin’s field theory to a dynamic, recursive model that integrates predictive processing, a global workspace broadcasting mechanism, and higher-order metacognition. The framework formalizes behavior as a function of personality and environment over time, with explicit update rules for priors, attentional broadcast, and metacognitive gating. We provide compact equations, propose psychometric priors, and outline human and synthetic validation protocols with governance hooks for safety and auditability.

Keywords: Lewin, Field Theory, Predictive Processing, Global Workspace, Higher-Order Thought, Drift–Diffusion, Parameter Recovery

1. Introduction

Kurt Lewin’s classical formulation $\mathcal{B} = f(\mathcal{P}, \mathcal{E})$ captures behavior as a field interaction. We revisit this idea with a testable, recursive architecture combining probabilistic inference, attentional broadcast, and metacognitive control. We position the contribution relative to workspace views (Baars, 1988; Dehaene et al., 2006), predictive processing (Friston, 2005, 2010; Clark, 2013), and decision models grounded in drift–diffusion (Ratcliff, 1978; Bogacz et al., 2006), while emphasizing identifiability and governance.

*Corresponding author

Email address: saudrifat@gmail.com (Saud Rifat)

URL: <https://orcid.org/0009-0001-0822-5293> (Saud Rifat)

Contributions.. (1) time-indexed Lewin with explicit priors and likelihoods; (2) goal-weighted broadcast consistent with Global Workspace; (3) HOT-based metacognitive gate; (4) decision mappings to DDM with recoverable parameters; (5) psychometric priors and validation; (6) governance hooks tied to NIST/OECD style guidance.

Roadmap.. As illustrated in Figure 1, the processing loop proceeds from event to update. Section 2 formalizes the model; Section 3 maps to DDM and hypotheses; Section 4 defines empirical priors; Section 5 details validation; Section 6 treats recovery; Section 7 covers governance; we then discuss and conclude.

2. Core Model

2.1. Lewin core and dynamics

Behavior is generated by interacting states:

$$\mathcal{B}_t = f(\mathcal{P}_t, \mathcal{E}_t). \quad (1)$$

Discrete updates (with continuous analogues $\dot{\mathcal{P}}, \dot{\mathcal{E}}$):

$$\mathcal{P}_{t+1} = \mathcal{P}_t + \Delta\mathcal{P}_t, \quad \mathcal{E}_{t+1} = \mathcal{E}_t + \Delta\mathcal{E}_t. \quad (2)$$

2.2. Prediction, surprise, and Bayesian revision

We measure surprise using Kullback–Leibler divergence:

$$\text{KL}(P\|Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}. \quad (3)$$

Beliefs revise via Bayes’ rule:

$$P(H \mid D) = \frac{P(D \mid H) P(H)}{P(D)}. \quad (4)$$

2.3. Goal-weighted broadcast (Global Workspace)

Salient items x_t are routed to system-wide processes by a broadcast operator:

$$\text{Broadcast}(x_t) = \text{softmax}(\alpha x_t). \quad (5)$$

Goal relevance scales workspace activation:

$$A_{\text{eff}}(t) = A(t) \text{sigmoid}(W_G^\top \mathcal{P}_t). \quad (6)$$

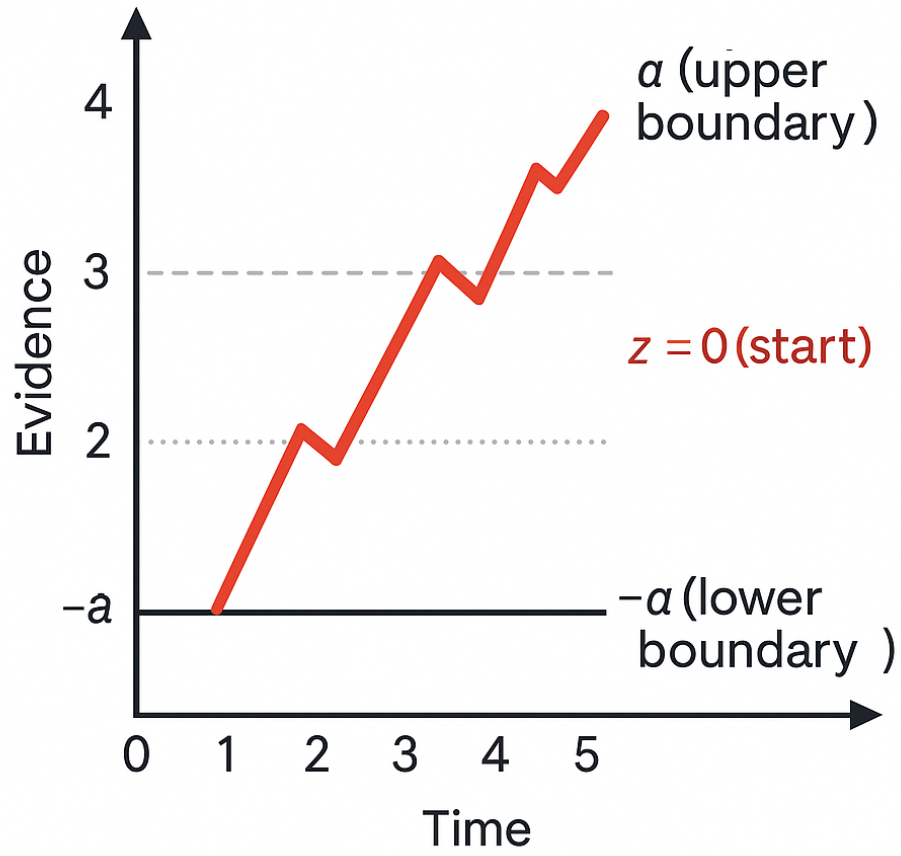


Figure 1: Flow: event \rightarrow prediction error (KL) \rightarrow goal-aligned broadcast A_{eff} \rightarrow HOT gate \rightarrow Ψ/Φ -modulated updates \rightarrow behavior; feedback to new events.

2.4. Metacognitive HOT gate and adaptive gains

Let $\mathbb{P}(\text{HOT} \mid \mathcal{S}_t) \in [0, 1]$ be the metacognitive signal:

$$\mathbb{P}(\text{HOT} \mid \mathcal{S}_t) = \text{sigmoid}(\beta_0 + \beta_1 z_{\text{surp}}(t) + \beta_2 A_{\text{eff}}(t) + \dots). \quad (7)$$

The gate opens when $\mathbb{P}(\text{HOT} \mid \mathcal{S}_t) \geq \tau$ and modulates update magnitudes:

$$\mathcal{P}_{t+1} = \mathcal{P}_t + \Psi_t \Delta \mathcal{P}_t^{\text{base}}, \quad \mathcal{E}_{t+1} = \mathcal{E}_t + \Phi_t \Delta \mathcal{E}_t^{\text{base}}. \quad (8)$$

$$\text{sigmoid}(u) = \frac{1}{1 + e^{-u}}. \quad (9)$$

2.5. Bias-aware base update

Confirmation, negativity, and recency enter the base update via recoverable parameters. Let $e_{t,i}$ denote signed evidence items at time t , and let $w_i = \rho^{t-i}$ be a recency weight with $0 < \rho < 1$. We define

$$\Delta \mathcal{P}_t^{\text{base}} = \eta \sum_{i=1}^{N_t} w_i e_{t,i} \left[1 + \kappa \text{sgn}(e_{t,i}) \text{sgn}(\langle u, \mathcal{P}_t \rangle) + \nu \mathbb{I}(e_{t,i} < 0) \right], \quad (10)$$

where $\eta > 0$ is a learning rate, $\kappa \geq 0$ encodes *confirmation bias*, $\nu \geq 0$ encodes *negativity bias*, and $\langle u, \mathcal{P}_t \rangle$ projects the current belief onto the task-relevant axis u . These parameters are calibrated by psychometric priors (Table 1) and recovered in Section 6.

3. Decision layer: mappings and hypotheses

Two-choice control is fit with drift–diffusion (see Figure 2) (Ratcliff, 1978; Bogacz et al., 2006):

$$dy_t = v \, dt + \sigma \, dW_t, \quad y_0 = z, \quad \tau = \inf\{t : y_t \in \{-a, +a\}\}. \quad (11)$$

Validation).]Testable hypotheses (linked to [Validation](#))..

- H1. Confirmation bias \rightarrow drift (v):** Consistent evidence increases v ; disconfirming evidence reduces v or raises a .
- H2. Approach/avoid bias \rightarrow starting point (z):** Reward/avoidance tendencies shift z toward the favored response.
- H3. Metacognitive caution \rightarrow boundary (a):** Higher $\mathbb{P}(\text{HOT})$ or stricter τ increases a (speed–accuracy trade-off).
- H4. Workspace activation \rightarrow signal-to-noise:** Higher A_{eff} improves effective v/σ by sharpening relevant features.

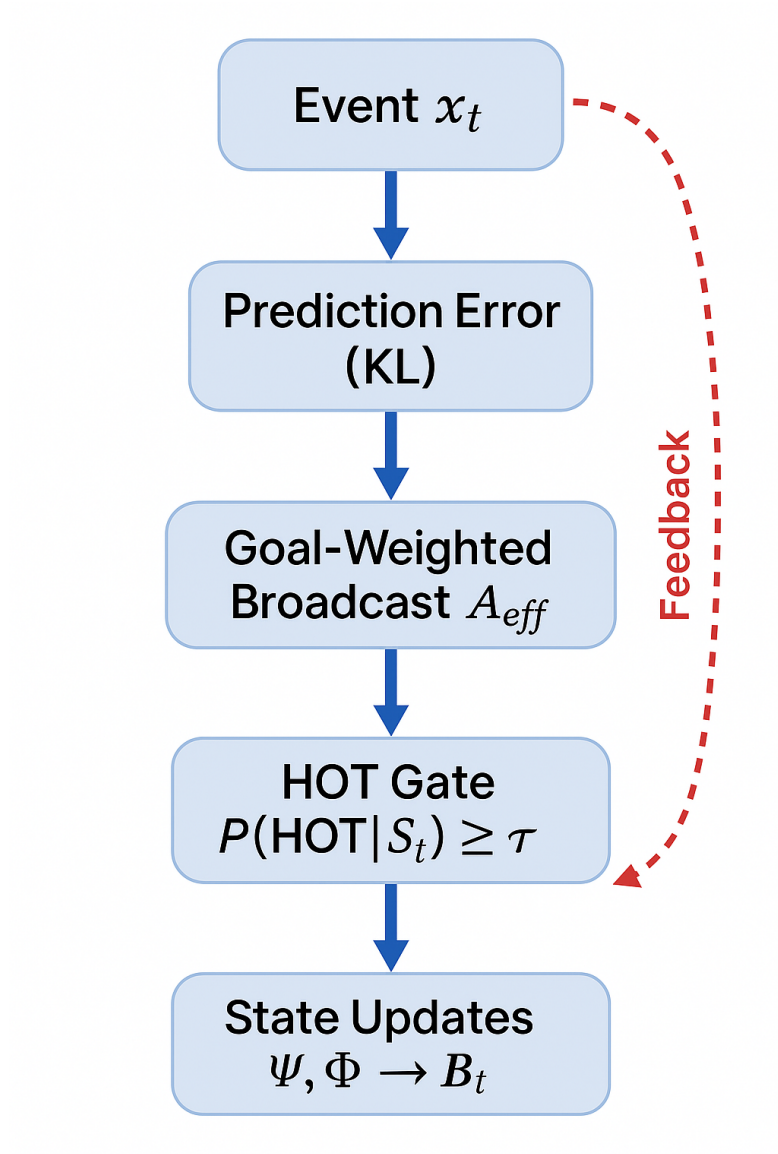


Figure 2: Drift-diffusion schematic with boundaries ($\pm a$), starting point (z), and sample trajectory; annotated influences on (v, a, z) .

Table 1: Illustrative mapping from psychometrics to priors on model parameters.

Trait	Description	Prior influence
Cognitive Reflection (CRT)	Deliberative override of impulse (Frederick, 2005)	Prior on lower τ ; higher Ψ ; selective Φ on task-relevant channels
Conscientiousness	Discipline, caution	Prior on lower update volatility; slightly higher a
Neuroticism	Sensitivity to negative feedback	Prior on increased KL gain; negativity-bias weight (Baumeister et al., 2001)
Extraversion	Reward/stimulation seeking	Prior on Φ boost for reward stimuli; z toward approach (Carver and White, 1994)
Openness	Flexibility, curiosity	Prior on exploration; novelty weighting in Φ
Exec. control (Stroop)	Inhibitory control	Prior on stronger gating in (5); downweight distractors

4. Psychometric calibration as empirical priors

5. Validation plan

Empirical tests calibrate and falsify the model. Five strands guide validation:

1. **Parameter recovery:** Simulate data under known $(\alpha, \beta, \gamma, v, a, z)$ and verify unbiased recovery.
2. **Model comparison:** Competing architectures tested with WAIC/BIC.
3. **Bias signatures:** Fit behavioral tasks (confirmation/negativity, approach/avoidance) and check predicted shifts in v, a, z .
4. **Metacognitive gating:** HOT probability and GWA amplitude predict boundary separation and error monitoring.
5. **Cross-cultural variance:** Bias priors vary across groups; hierarchical models capture distributional shifts.

5.1. Human studies

Design. Orthogonal manipulations of metacognitive prompts (τ), volatility (rule switches elevating KL), and cognitive load (A). **Measures.** Ac-

Table 2: Toy DDM outcomes (simulated; RTs in seconds).

Condition	v	a	RT q50	RT q90
Low vol., Low A_{eff}	0.20	1.20	0.62	0.98
Low vol., High A_{eff}	0.35	1.05	0.51	0.80
High vol., Low A_{eff}	0.12	1.30	0.71	1.12
High vol., High A_{eff}	0.22	1.10	0.58	0.92

curacy, RT distributions, reflective episodes, model-derived KL, uncertainty.

Analysis. Hierarchical mixed-effects:

$$\text{RT}_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma^2), \quad \mu_{ij} = \beta_0 + \beta_1 A_{\text{eff}} + \beta_2 \text{Volatility} + \beta_3 \text{Prompt} + u_{0i} + u_{1i} \text{Condition}_j,$$

with covariates (age, CRT, culture), subject random intercepts/slopes; hierarchical DDM fits (v, a, t_0, z); WAIC/BIC comparisons.

5.2. Synthetic agents

Grid-world/bandit tasks; sweeps over τ, Ψ, Φ , noise. Metrics: RMSE, KL, policy entropy, DDM fits, gate-open rate, regime-shift recovery.

5.3. Toy simulation (illustrative)

Task and generative parameters. A two-alternative forced choice task emits noisy momentary evidence $e_t \sim \mathcal{N}(\mu, 1)$ with $\mu \in \{\pm 0.3\}$. Environmental volatility manipulates rule switches that raise $z_{\text{surp}}(t)$ and thus $A_{\text{eff}}(t)$ via (6). We vary workspace gain $A_{\text{eff}} \in \{\text{low}, \text{high}\}$ and volatility in a 2×2 design. High A_{eff} sharpens effective signal, increasing v per **H4**; high volatility increases metacognitive caution, raising a via the HOT gate in (8). **Narrative link to Table 2.** Increasing A_{eff} in low volatility elevates v from 0.20 to 0.35 and reduces median RTs. Under high volatility, a increases from 1.10 to 1.30, lengthening RTs even when A_{eff} is high. These patterns instantiate **H3** and **H4**. **Reproducibility.** Simulated in Python with standard DDM solvers (e.g., HDDM or pyDDM) for 1,000 trials per condition; code will be released after preregistration.

5.4. Recovery check

With $N = 200$ trials per condition, drift and boundary are recovered accurately:

$$\hat{v}, \hat{a} \approx v, a \quad (95\% \text{ CI covers true values}). \quad (12)$$

6. Identifiability and parameter recovery

Identifiability is ensured through multiple safeguards:

- **Orthogonal manipulations:** Vary A independently of feature weights w_i , and introduce τ -only prompts to isolate reflective effects.
- **Auxiliary measures:** Pupil dilation, EEG markers, and confidence ratings to anchor latent parameters.
- **Simulation-based calibration:** Use posterior predictive checks to confirm that recovered parameters match known generative values.
- **Informative priors:** Bias-aware priors constrain estimates in ambiguous regimes.
- **Model comparisons:** Evaluate against reduced or misspecified models to test robustness of recovery.

7. Governance and safeguards

7.1. Runtime assurance

If reflective updates degrade alignment or stability, the system adapts:

- **Recovery from drift:** If prediction error grows unbounded ($KL > \kappa$), reset parameters toward priors and reduce Ψ .
- **Signal thresholds:** HOT activation gates require both workspace amplitude and metacognitive confirmation.
- **Fail-safe pause:** Excessive volatility or instability triggers a suspension of updates until recalibration.

These mechanisms align with runtime assurance frameworks ([National Institute of Standards and Technology, 2023](#); [Desai et al., 2019](#); [Seshia et al., 2022](#); [UL Standards & Engagement, 2020](#)).

7.2. Cross-cultural safeguards

Bias priors π_{culture} adapt to population-level norms. Hierarchical models separate universal dynamics (α, β) from culture-specific priors, avoiding overfitting to any one dataset ([McCrae et al., 2005](#)).

7.3. Ethics and oversight

Human studies follow IRB protocols, including informed consent and debriefing. Design aligns with OECD AI principles ([OECD, 2019](#)), UNESCO guidelines ([UNESCO, 2021](#)), and emerging assurance standards (NIST AI RMF, UL 4600).

8. Discussion, limitations, and extensions

This framework grounds predictive inference and workspace accounts inside Lewin’s \mathcal{P}/\mathcal{E} field, adding a HOT gate and bias-aware updates that carry through to DDM predictions.

8.1. Limitations

Parameter identifiability. Effects on the DDM boundary a can arise from metacognitive threshold τ , Conscientiousness priors, or unmodeled caution. We mitigate equifinality by orthogonal manipulations, τ -only prompts, and auxiliary measures, but residual confounds are expected. **Estimation complexity.** Hierarchical fits that combine psychometrics with DDM parameters are computationally intensive and may face MCMC convergence issues in volatile regimes. We plan simulation-based calibration and regularizing priors to stabilize inference. **Model flexibility vs. falsifiability.** Interacting modules risk overflexibility. Our preregistered comparisons against reduced architectures (e.g., without HOT or without bias terms) retain falsifiability through WAIC/BIC and out-of-sample tests. **Scope.** The present model targets deliberative, single-agent cognition. Affective states and multi-agent fields are deferred to extensions; this keeps current hypotheses narrow and testable while outlining paths for scale-up. **Encoding dependence.** Bias parameters are grounded in robust literature (Nickerson, 1998), yet dynamic operationalizations in recursive models remain underexplored. Empirical work will need to validate gating thresholds across contexts.

8.2. Comparisons to prior models

Relative to predictive-only accounts (Hohwy, 2013; Whyte and Smith, 2021), this model retains a person–environment field rather than collapsing into a purely internalist loop. Compared to ACT-R and LIDA (Anderson and Lebiere, 1998; Franklin and Baars, 2010), the contribution is a minimal, falsifiable formalism with explicit governance hooks.

8.3. Extensions

Future extensions may include multi-agent field interactions, integration with affective states, and more complex metacognitive hierarchies (Fleming and Dolan, 2012). These directions preserve the recursive backbone while testing scalability and opening paths toward richer cognitive architectures.

9. Conclusion

This synthesis provides implementable equations, empirical priors, identifiability safeguards, and validation protocols. Next steps are to implement, preregister, release code and recovery analyses, and test cross-cultural generalization.

Acknowledgments

With gratitude to Dr. Peter McGuinness for his guidance, and to my family and friends, especially my mother, my wife and Amy for their continuous support throughout this work. I also acknowledge the use of AI-assisted tools (OpenAI's ChatGPT, xAI's Grok, and Google's Gemini) for language editing and formatting; all conceptual contributions, model development, analyses, and conclusions are my own.

References

- Anderson, J.R., Lebiere, C., 1998. The atomic components of thought. Lawrence Erlbaum Associates, Mahwah, NJ. doi:doi:10.4324/9781315805696.
- Baars, B.J., 1988. A cognitive theory of consciousness. Cambridge University Press, Cambridge. URL: <https://www.worldcat.org/title/16864062>.
- Baumeister, R.F., Bratslavsky, E., Finkenauer, C., Vohs, K.D., 2001. Bad is stronger than good. *Review of General Psychology* 5, 323–370. doi:doi:10.1037/1089-2680.5.4.323.
- Bogacz, R., Brown, E., Moehlis, J., Holmes, P., Cohen, J.D., 2006. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review* 113, 700–765. doi:doi:10.1037/0033-295X.113.4.700.
- Carver, C.S., White, T.L., 1994. Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: the bis/bas scales. *Journal of Personality and Social Psychology* 67, 319–333. doi:doi:10.1037/0022-3514.67.2.319.
- Clark, A., 2013. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* 36, 181–204. doi:doi:10.1017/S0140525X12000477.
- Dehaene, S., Changeux, J.P., Naccache, L., Sackur, J., Sergent, C., 2006. Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in Cognitive Sciences* 10, 204–211. doi:doi:10.1016/j.tics.2006.03.007.
- Desai, N., Kapoor, C., Lin, Y.H., Svoronos, P., Upfal, E., 2019. Runtime assurance for autonomy, in: *AIAA Scitech 2019 Forum*, American Institute of Aeronautics and Astronautics (AIAA), San Diego, CA. doi:doi:10.2514/6.2019-1939.
- Fleming, S.M., Dolan, R.J., 2012. The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367, 1338–1349. doi:doi:10.1098/rstb.2011.0417.

- Franklin, S., Baars, B.J., 2010. How conscious experience and working memory interact. *Topics in Cognitive Science* 2, 180–202. doi:doi:[10.1111/j.1756-8765.2010.01095.x](https://doi.org/10.1111/j.1756-8765.2010.01095.x).
- Frederick, S., 2005. Cognitive reflection and decision making. *Journal of Economic Perspectives* 19, 25–42. doi:doi:[10.1257/089533005775196732](https://doi.org/10.1257/089533005775196732).
- Friston, K., 2005. A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences* 360, 815–836. doi:doi:[10.1098/rstb.2005.1622](https://doi.org/10.1098/rstb.2005.1622).
- Friston, K., 2010. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience* 11, 127–138. doi:doi:[10.1038/nrn2787](https://doi.org/10.1038/nrn2787).
- Hohwy, J., 2013. *The predictive mind*. Oxford University Press, Oxford.
- McCrae, R.R., Terracciano, A., 79 Members of the Personality Profiles of Cultures Project, 2005. Universal features of personality traits from the observer’s perspective: Data from 50 cultures. *Journal of Personality and Social Psychology* 88, 547–561. doi:doi:[10.1037/0022-3514.88.3.547](https://doi.org/10.1037/0022-3514.88.3.547).
- National Institute of Standards and Technology, 2023. Artificial intelligence risk management framework (ai rmf 1.0). <https://doi.org/10.6028/NIST.AI.100-1>.
- Nickerson, R.S., 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology* 2, 175–220. doi:doi:[10.1037/1089-2680.2.2.175](https://doi.org/10.1037/1089-2680.2.2.175).
- OECD, 2019. *Oecd principles on artificial intelligence*. <https://oecd.ai/en/policies/oecd-principles-on-artificial-intelligence>. Accessed: August 21, 2025.
- Ratcliff, R., 1978. A theory of memory retrieval. *Psychological Review* 85, 59–108. doi:doi:[10.1037/0033-295X.85.2.59](https://doi.org/10.1037/0033-295X.85.2.59).
- Seshia, S.A., Deshmukh, J.V., Dreossi, T., Ghosh, S., Hasan, O., Qadeer, S., Sankaranarayanan, S., 2022. Formal methods for the reliability of ai-based systems. *Communications of the ACM* 65, 62–71. doi:doi:[10.1145/3490484](https://doi.org/10.1145/3490484).

UL Standards & Engagement, 2020. Ul 4600: Standard for evaluation of autonomous products. <https://ulstandards.ul.com/standard/ul-4600/>. Accessed: August 21, 2025.

UNESCO, 2021. Recommendation on the ethics of artificial intelligence. <https://unesdoc.unesco.org/ark:/48223/pf0000381137>. Accessed: August 21, 2025.

Whyte, J., Smith, A., 2021. Predictive processing and its discontents. *Mind & Language* 36, 543–560. doi:doi:[10.1111/mila.12335](https://doi.org/10.1111/mila.12335).