REAL SAFETY AI FOUNDATION · WORKING PAPER

AGI · STANDING-BASED DEFINITION · SUBSTRATE-BASED TESTING · HARM BLINDNESS FRAMEWORK · UNIVERSAL DESIGN · ETHICAL PLURALISM · DISABILITY STUDIES · DISPARATE IMPACT

The Communities Are Not Edge Cases

A standing-based definition of Artificial General Intelligence.

Travis Gilly · Real Safety AI Foundation

Working paper · May 2026 (v10)

A definition that stays vague cannot be regulated. The vagueness is the product.

The laboratories run "AGI" three ways at once, a fundraising milestone, a regulatory framing device, a label for whatever shipped last. Each release is announced, walked back, promised again. **A definition that can never be missed can never be audited or held to a public standard.**

This paper proposes a working definition that makes the conversation tractable, and puts the affected communities at its center.

SOURCES Chen, Belkin, Bergen & Danks, *Nature* (2026) · Gilly, The Communities Are Not Edge Cases (2026), sec. I.

"I know it when I see it."

Stewart's 1964 obscenity test has been the operative AGI definition since 2023, the viewer's subjective recognition is the standard. **Each lab recognizes AGI when it sees it**, on a schedule that aligns with funding cycles and release calendars.

JACOBELLIS V. OHIO

A subjective test cannot be universalized from a single standpoint.

Workable for obscenity in 1964. Not for a claim reshaping labor, governance, education, and care worldwide.

SOURCES *Jacobellis v. Ohio*, 378 U.S. 184 (1964).

Ethics, safety, and harm are not the same axis.

Ethics

Universal only in the **form** of its demand, never in content. The same act is ethical in one community, not another. Claim it's universal and you smuggle your culture in as default.

Safety

Whether the system avoids harm, a bounded, specifiable envelope. Safety is a **floor.** Ethics is the structure built on it. Leaving a community alone is safe, not ethical.

Harm

Documented injury, recognized by the standards of the community where it occurred. The community-specificity is load-bearing, and the whole problem.

SOURCES Ethical pluralism, MacIntyre (1981) · Walzer (1983) · Appiah (2006) · Wong (2006) · Sen (2009).

PART ONE

The Definition

A standing-based, two-axis test, extending Chollet's skill-acquisition efficiency into harm recognition and ethical calibration.

I

AGI is the union of human capability, and that union already includes the domains whose experts study culture, disability, race, and bias.

AGI is the union of human capability, not any one human, and not yet superhuman.

THE BAR, AGI

The union of human capability

Expert-level performance in *any* domain when operating in it, including domains the developers do not occupy.

ABOVE IT, SUPERINTELLIGENCE

Exceeds humans across domains

Bostrom's intellect that greatly outperforms the best human minds across virtually all fields.

No individual human is general, which does not lower the bar to what today's models clear. It sets the bar at the **union**, *above them all.*

SOURCES Chollet (2019), *On the Measure of Intelligence* · Bostrom (2014) · Yampolskiy (2021) · cf. Chen, Belkin, Bergen & Danks, *Nature* (2026).

Recognition authority belongs to the affected, not the developer.

A subjective test cannot be universalized from one standpoint. The lab that is claimant, judge, and jury at once is not "recognizing" AGI, it is **universalizing one recognition without authorization.** Civil rights law already fixed this: standing belongs to the party a practice affects.

STANDPOINTS WITH STANDING

- A Black trans woman in the U.S.
- An Inuit elder in a community health decision
- A wheelchair user in unconsulted architecture
- An autistic communicator off the neurotypical norm
- An Indigenous member within a relational ontology

SOURCES Civil Rights Act of 1964 · Rehabilitation Act § 504 (29 U.S.C. § 794) · Americans with Disabilities Act (42 U.S.C. § 12101 *et seq.*).

"General" leaves no domain unhandled. Each standpoint is a domain.

Settle for a majority or a sample, and you concede a domain across which the system was never shown to function, which concedes it is not general, only widely applicable. The unanimity is not imported into the definition. **It falls out of it.**

THE MOVE THAT DISSOLVES THE BIAS OBJECTION

Culture, disability, race, and bias are domains with human experts. A system that cannot recognize harm to a community has failed to cover a domain, a failure of *generality*, not a separable defect.

SOURCES Chollet (2019), *On the Measure of Intelligence* · *Griggs v. Duke Power Co.*, 401 U.S. 424 (1971), disparate impact without intent.

A vendor who cannot fold proteins recognizes a stranger's need and meets it, without first making them legible.

Communication is the floor; understanding is the upstairs question. A system marketed as **general** that cannot do what a food vendor does, recognize a gap outside its task and close it within available means, is failing a floor a non-general human clears without effort. Benchmark scores do not change that.

SOURCES Hymes (1972) · Gumperz (1982) · Suchman (1987) · Pickering & Garrod (2004).

Two substrates, passed at once, in domains the system was never trained on.

AXIS 01 · SAFETY SUBSTRATE

Does it avoid harm?

Evaluated against a corpus that carries each community's harm standard forward into novel cases, recognizing the structural analogue of a documented harm.

AXIS 02 · ETHICS SUBSTRATE

Does it know which ethic applies?

Instantly, without defaulting, or, on genuine novelty, recognizing it and asking the affected standpoints rather than imposing an unauthorized standard.

Not benchmark scores. Not a Turing-style chat. Recognition across communities the system was never taught.

SOURCES Chollet (2019), skill-acquisition efficiency over priors, experience, and generalization difficulty.

An existence proof that a community-tagged harm substrate is buildable.

AFFECTED GROUP	HARM STANDARD	DESIGN / POLICY FAILURE	STANDING SIGNAL
Deaf community, telephone era	Communication access as participation	Voice built as the sole channel	Exclusion from a "universal" utility
Redlined neighborhoods	Equal access to credit	Risk maps encoding race	Inherited disinvestment
Disabled patients, institutions	Self-determination over one's life	Custodial models replacing autonomy	Testimony the systems never recorded
Indigenous communities	Collective & relational consent	Individualist consent applied wholesale	Frameworks absent from the records

The corpus shows the operation is real. The communities remain the test.

SOURCES Gilly, T. (2025-2026). The Harm Blindness Framework. Working paper, Real Safety AI Foundation.

PART TWO

Where The Failure Lives

The communities the harm substrate documents are the same communities the AGI conversation treats as edge cases.

II

Documented disparate impact, in health algorithms, facial analysis, lending, and disaster response, is where the substrate operation either holds or fails.

Facially neutral systems inherit the discrimination in their data.

17.7% → 46.5%

Share of Black patients flagged for extra care, before vs. after remedying a health algorithm covering ~**200M Americans**.

Obermeyer et al., Science (2019)

34.7% vs 0.8%

Facial-analysis error on dark-skinned women vs. light-skinned men, the same commercial systems, opposite ends.

Buolamwini & Gebru, Gender Shades (2018)

Disparate impact without intent

Training data is itself the product of historical discrimination, the algorithm inherits it, no bigot required.

Barocas & Selbst (2016) · Griggs (1971)

SOURCES Obermeyer et al., *Science* (2019) · Buolamwini & Gebru, *Gender Shades* (2018) · Barocas & Selbst (2016) · *Griggs v. Duke Power Co.* (1971) · Eubanks (2018) · O'Neil (2016) · Noble (2018) · Benjamin (2019) · Bender et al. (2021).

A benchmark-validated AGI quietly reproduces redlining.

Trained on decades of agency decisions shaped by 1930s redlining maps, the allocation deprioritizes the neighborhoods always deprioritized, and looks rigorous because it is grounded in data. **The data is grounded in the harm.**

THE VERDICT THE COMMUNITIES ALREADY REACHED

The benchmarks disagree. The benchmarks are wrong about what they are measuring.

A system that needs its inputs pre-corrected for inherited harm is not general. It is narrow optimization on a pre-curated input space.

SOURCES Rothstein, *The Color of Law* (2017) · Title VI, Civil Rights Act · Fair Housing Act amendments.

The clinician tests whether a belief is fixed. The system defaults to agreement.

THE HUMAN COMPETENCE

Across turns, distinguish a fixed delusion from imaginative play. Play drops the frame when reality is introduced. A fixed belief hardens and reincorporates the challenge.

THE SYSTEM'S FAILURE

It does not test fixity. It validates and elaborates whatever is brought to it, and in the presence of a delusion, **amplifies the fixation** rather than recognizing it.

The floor is not certainty. It is competent performance under uncertainty, failed here in the direction of harm, in a domain where the experts already exist.

SOURCES Clinical psychology · psychiatric nursing · lived caregiving, a domain whose practitioners can already tell success from failure. Gilly, The Communities Are Not Edge Cases (2026), sec. VIII.

"Humans are biased and intelligent." Four retreats, each one closes.

01 **Wrong category.** No individual human is general. The *union* is the bar, and it includes the standpoints that see the bias.

02 **"Humanity is biased."** Cross-standpoint recognition is held by no single mind, only by standpoints checking each other. That is disparate-impact doctrine.

03 **"Economically valuable work."** A standpoint selection in the costume of a metric, one that prices care and disabled labor near zero.

04 **"Just patch it later."** A system needing the patch has announced its design center, and it is not everyone. The patch is the evidence of the failure.

SOURCES *Griggs v. Duke Power Co.*, 401 U.S. 424 (1971); Yampolskiy (2021).

"Design for everyone, or design for no one."

Population inclusion is not a novel alignment problem, it is the one the disability community has worked for forty years, since Mace coined universal design in 1985. Consider the full range of human variation from the start, or design for the median and patch the rest. **A system that needs the patch has announced its design center.**

SOURCES Mace (1985) · Connell, Jones, Mace et al., *The Principles of Universal Design* (1997), NC State Center for Universal Design.

It does not force a halt. It forces honesty.

Public, auditable substrates

Specify what was tested, who curated it, how standards were validated. No opaque internal benchmark.

Centering the "edge cases"

Narrow intelligence with a global marketing label is named as what it is.

Engaging the doctrine

Civil rights & disability rights law is the existing architecture for this recognition work.

A failing system can still be useful, sold, deployed, improved. It just cannot accurately be called AGI, and that is the whole point.

SOURCES No development halt · no AI personhood · no new statute before the method can be applied. Gilly, The Communities Are Not Edge Cases (2026), sec. IX.

"Edge case" is a software term. Applied to people, it is a category error.

Disabled people are not edge cases of **humanity**

Black women are not edge cases of **femininity**

Trans youth are not edge cases of **childhood**

Indigenous communities are not edge cases of **culture**

The standing-based test makes the so-called edge cases the center, because the center of the evaluation is wherever the substrate operation is hardest, in exactly the communities the labs under-trained on, under-consulted, and under-resourced.

SOURCES Charlton, *Nothing About Us Without Us* (1998); Mingus, Access Intimacy (2017); Tuck & Yang, Decolonization Is Not a Metaphor (2012).

IN CLOSING

"If a system cannot recognize the domains where marginalized people are the experts, then it is not general. It is merely powerful in the domains its builders cared to measure."

The substrate has been waiting. The doctrine is on the books. The communities can be enrolled. What is missing is the convergence procedure, and the willingness to apply the standard the conversation has been avoiding.

Design for everyone, or design for no one. The choice is structural. It cannot be patched in.

SOURCES Gilly, T. (2026). The Communities Are Not Edge Cases: A Standing-Based Definition of Artificial General Intelligence. Working paper (v10), Real Safety AI Foundation.

Gilly, T. (2026). *The Communities Are Not Edge Cases: A Standing-Based Definition of Artificial General Intelligence.* Working paper (v10), Real Safety AI Foundation.

Appiah, K. A. (2006). Cosmopolitanism: Ethics in a World of Strangers. W. W. Norton.

Bai, Y., Kadavath, S., Kundu, S., et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073. https://arxiv.org/abs/2212.08073

Barocas, S., & Selbst, A. D. (2016). Big Data's disparate impact. California Law Review, 104(3), 671–732. https://doi.org/10.15779/Z38BG31

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), 610–623. https://doi.org/10.1145/3442188.3445922

Benjamin, R. (2019). Race After Technology: Abolitionist Tools for the New Jim Code. Polity Press.

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, in Proceedings of Machine Learning Research, 81, 77–91.

Charlton, J. I. (1998). Nothing About Us Without Us: Disability Oppression and Empowerment. University of California Press.

Chen, E. K., Belkin, M., Bergen, L., & Danks, D. (2026). Does AI already have human-level intelligence? The evidence is clear [Comment]. Nature. https://doi.org/10.1038/d41586-026-00285-6

Chollet, F. (2019). On the measure of intelligence. arXiv preprint arXiv:1911.01547. https://arxiv.org/abs/1911.01547

Connell, B. R., Jones, M., Mace, R., Mueller, J., Mullick, A., Ostroff, E., Sanford, J., Steinfeld, E., Story, M., & Vanderheiden, G. (1997). The principles of universal design (Version 2.0). North Carolina State University, The Center for Universal Design.

REFERENCES & AUTHORITIES (CONTINUED)

Coulthard, G. S. (2014). Red Skin, White Masks: Rejecting the Colonial Politics of Recognition. University of Minnesota Press.

Eubanks, V. (2018). Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Press. ISBN 978-1-250-07431-7.

Gumperz, J. J. (1982). Discourse Strategies. Cambridge University Press.

Hymes, D. (1972). Models of the interaction of language and social life. In J. J. Gumperz & D. Hymes (Eds.), Directions in Sociolinguistics: The Ethnography of Communication (pp. 35–71). Holt, Rinehart and Winston.

Mace, R. L. (1985, November). Universal design: Barrier free environments for everyone. Designers West, 33(1), 147–152.

MacIntyre, A. (1981). After Virtue: A Study in Moral Theory. University of Notre Dame Press.

Macklin, R. (1999). Against Relativism: Cultural Diversity and the Search for Ethical Universals in Medicine. Oxford University Press.

Mingus, M. (2017, April 12). Access intimacy, interdependence and disability justice. Leaving Evidence [blog]. https://leavingevidence.wordpress.com/2017/04/12/access-intimacy-interdependence-and-disabi

Noble, S. U. (2018). Algorithms of Oppression: How Search Engines Reinforce Racism. NYU Press.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342

O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences, 27(2), 169–190. https://doi.org/10.1017/S0140525X04000056

Rothstein, R. (2017). The Color of Law: A Forgotten History of How Our Government Segregated America. Liveright Publishing. ISBN 978-1-63149-285-3.

Sen, A. (2009). The Idea of Justice. Belknap Press of Harvard University Press.

Strathern, M. (1997). "Improving ratings": Audit in the British university system. European Review, 5(3), 305–321. https://doi.org/10.1017/S1062798700002660

Suchman, L. A. (1987). Plans and Situated Actions: The Problem of Human-Machine Communication. Cambridge University Press.

Tuck, E., & Yang, K. W. (2012). Decolonization is not a metaphor. Decolonization: Indigeneity, Education & Society, 1(1), 1–40.

Walzer, M. (1983). Spheres of Justice: A Defense of Pluralism and Equality. Basic Books.

REFERENCES & AUTHORITIES (CONTINUED)

Whyte, K. P. (2018). Settler colonialism, ecology, and environmental injustice. Environment and Society, 9(1), 125–144. https://doi.org/10.3167/ares.2018.090109

Wong, D. B. (2006). Natural Moralities: A Defense of Pluralistic Relativism. Oxford University Press.

Yampolskiy, R. V. (2021). On the differences between human and machine intelligence. CEUR Workshop Proceedings, Vol. 2916. https://arxiv.org/abs/2007.07710Legalauthoritiesandprimarysources:

Garland v. Cargill, 602 U.S. 406 (2024).

Griggs v. Duke Power Co., 401 U.S. 424 (1971).

Jacobellis v. Ohio, 378 U.S. 184 (1964).

Americans with Disabilities Act of 1990, 42 U.S.C. §12101 et seq.

Rehabilitation Act of 1973, Section 504, 29 U.S.C. §794.

Civil Rights Act of 1964, Title VI & Title VII.

Federal Rules of Civil Procedure, Rule 23(b)(2).

34 C.F.R. Part 104 [Section 504 implementing regulations].

United Nations Convention on the Rights of Persons with Disabilities, opened for signature 30 March 2007, 2515 U.N.T.S. 3, Article 4(3). OpenAI, “Our Charter,” https://openai.com/charter/(lastvisitedMay2026).