The Communities Are Not Edge Cases:

A Standing-Based Definition of Artificial General Intelligence

Travis Gilly
ORCID: 0009-0007-2954-6313
Real Safety AI Foundation
Illinois, United States

Working paper, May 2026 (v10)

ABSTRACT

The contemporary conversation about Artificial General Intelligence operates without a working definition. Major laboratories use the term as a fundraising milestone, a regulatory framing device, and a release marker for whatever capability the laboratory has most recently shipped. The resulting ambiguity is functionally a control mechanism: a definition that stays vague cannot be regulated, audited, or held to a public standard. This paper proposes a standing-based, two-axis definition of AGI that extends Chollet's (2019) skill-acquisition-efficiency framework into the domains of harm recognition and ethical calibration. The definition treats AGI as the union of human capability, the ability to do what any human can do at expert level when operating in a domain, distinct from superintelligence, which exceeds human capability across domains. Because that union includes the domains whose experts study culture, disability, race, and bias, recognizing harm to a community is not an add-on to intelligence but one of the domains the definition already covers. AGI is recognized at the threshold where a system can be applied to any domain without the human first having to teach it that domain, evaluated against substrates that carry community-specific harm standards and ethical frameworks, with recognition authority allocated to the affected standing vantage points rather than to the developer. The Harm Blindness Framework, a 312-case corpus of historical and corporate harms across cultures and centuries, is offered as an existence proof that such a harm substrate is buildable and operable, not as a complete one. The ethics substrate is the next required build. The paper grounds the two-axis definition in the civil rights and disability rights tradition, the universal design methodology developed by Mace and colleagues at North Carolina State University, and the philosophy of language tradition's distinction between communication and understanding. The communities the AGI conversation treats as edge cases are placed at the center of the evaluation, because the center of the evaluation is wherever the recognition operation is hardest. The definition forces public, auditable substrates; forces engagement with civil rights and disability rights doctrine; and forces a structural choice for design that the disability community formalized four decades ago: design for everyone or design for no one.

Keywords: AGI, standing-based definition, substrate-based testing, Harm Blindness Framework, universal design, ethical pluralism, philosophy of language, communication, disability studies, disability rights, civil rights, disparate impact

I. Introduction

The conversation about Artificial General Intelligence in 2026 is operating with no working definition. In practice the major laboratories run the word three ways at once: as a fundraising milestone, as a device for framing the regulation that has not yet arrived, and as a label for whatever capability has most recently shipped. Each release is announced as AGI, walked back, then promised again. The ambiguity is the product. A definition that stays vague is a definition the laboratories can never be shown to have missed, and a definition that can never be missed can never be regulated, audited, or held to a public standard.[1]

Justice Potter Stewart's framing of obscenity in *Jacobellis v. Ohio*, 378 U.S. 184 (1964), captured by the line "I know it when I see it," has been functioning as the operative AGI definition since at least 2023. It worked, after a fashion, for obscenity in 1964. It is not adequate for a capability claim reshaping labor, governance, surveillance, education, healthcare and care of disabled, elderly and child populations across the world.

This paper proposes a standing-based, two-axis definition of AGI, grounded in working infrastructure that already exists and demonstrates the operation, and extends a definitional framework that Chollet (2019) has already articulated formally. Chollet defines intelligence as "skill-acquisition efficiency" with respect to "priors, experience, and generalization difficulty," such that a system is intelligent to the extent that it "is able to adapt to new problems it has not seen before and that its creators (developers) did not anticipate" (Chollet, 2019). The standing-based definition proposed here extends Chollet's framework into the domains of harm recognition and ethical calibration. The argument requires getting three preliminary terms straight first, because each of them is contested in ways the AGI conversation tends to flatten.

II. Three Terms the AGI Conversation Conflates

A. Ethics

Ethics is not universal in content. It is universal only in the form of the demand: every situation has a right answer that respects the parties involved, and the actor in the situation is responsible for finding it. The content of "right" varies by culture, community, history, relationship and stakes.

This is the philosophical position of ethical pluralism, which has been developed in moral philosophy from multiple directions over the past four decades (MacIntyre, 1981; Walzer, 1983; Appiah, 2006; Wong, 2006; Sen, 2009). A Western bioethics framework that treats individual patient autonomy as the highest value will misread a community where serious medical decisions are made collectively by family or by elders (Macklin, 1999). A United States disability rights framework that treats independent living as the goal will misread cultures where interdependence is the social ideal and where pursuing independence is read as a failure of care (Mingus, 2017). A liberal political philosophy that treats the autonomous individual as the unit of moral consideration will misread Indigenous frameworks where the unit of consideration is the relationship between persons, between persons and land, and between generations (Coulthard, 2014; Tuck & Yang, 2012; Whyte, 2018).

The same act can be ethical in one community and unethical in another, and the difference is not a matter of preference or accommodation. It is a matter of what the parties in the situation actually owe one another.

Anyone who tells you ethics is universal in content is either talking about meta-ethics, which is the universal form of the demand, or they are smuggling their own culture in as the default. The conflation produces the harm. A system

[1] For one major laboratory's own framing of AGI as the development of "highly autonomous systems that outperform humans at most economically valuable work," see OpenAI's published charter and mission statements at https://openai.com/charter/.

trained to respond ethically by a corporation in a single jurisdiction, on training data that overrepresents one culture's ethical assumptions, will apply that culture's ethic globally and call the result ethical AI. It is not. It is one culture's ethic with global reach.

B. Safety

Safety is the related but distinct question of whether the system avoids harm. Safety can be specified more narrowly than ethics because the set of unsafe outputs, while large, is bounded. A system that does not produce instructions for synthesizing biological weapons, does not generate sexual content involving children, does not give specific medical advice that would harm the user, does not output methods of self-harm, is operating within a recognized safety envelope. The envelope is contested at the edges and varies by jurisdiction, but the core is stable enough to be specified and tested.

Safety and ethics are not the same axis. A system can be safe to a marginalized community by simply leaving it alone, refusing to engage. That refusal is not ethical engagement. It is the same architectural failure that produces every "we didn't think about that population" disclosure after the fact. Safety is a floor; ethics is the structure built on the floor. A system that meets safety standards but cannot ethically engage with the communities it acts on is not ethical AI. It is sanitized AI, which is a different and lower achievement.

C. Harm

Harm is the term that has to be defined most carefully because it is the substrate the rest of the architecture sits on. Harm is not subjective preference. Harm is not what any individual claims to have suffered. Harm is documented injury, whether physical, psychological, social, economic or structural, recognizable to a trained observer applying the standards of the community in which the injury occurred.

The community-specificity is load bearing. A medical practice considered routine in one community can be harm in another. A communication style considered respectful in one culture can be insulting in another. A pedagogical method considered effective in one classroom can produce documented psychological injury in another. The harm is real in each case. The standards by which it is recognized vary.

This is the architectural problem AGI has to solve and the conversation has not yet acknowledged. The system has to recognize harm without flattening it to a single cultural standard and without dismissing community-specific harms as edge cases or as preferences. It has to know, instantly and without defaulting, which community is present and which standards apply.

III. The Harm Blindness Framework as Working Safety Substrate

The author developed the Harm Blindness Framework (HBF) over 2025 and 2026 to address exactly this problem at the design and policy levels.[2] The framework runs candidate designs, policies or products over a corpus of 312 documented historical and corporate harm cases drawn from across cultures, communities and centuries. Each case in the corpus is tagged with the community in which the harm occurred, the standards of harm recognition that community used, the design or policy decision that produced or failed to prevent the harm and the documented consequences. The cases

[2]The author has considered renaming the framework given the metaphorical use of blindness. One disability studies scholar with lived experience as a legally blind academic, asked directly, found the name uncomplicated in context. The author finds the intent-based defense uncomfortable on methodological grounds, since the framework itself rejects intent as a defense to documented harm. The name is in active reconsideration. The current working name will be revisited in a future paper. Any reader who finds the name produces harm in their context is invited to write to the author directly. The framework's own methodology will govern the response.

were assembled inductively from the historical record and from public corporate incident documentation, not deduced from a theoretical framework imposed on the cases.

A small sample conveys the structure of the corpus better than a description. The following rows are illustrative of the tagging scheme, not the full corpus:

Affected group	**Harm standard**	**Design or policy failure**	**Standing signal**
Deaf community, telephone era	Communication access as participation	Network built around voice as the sole channel	Exclusion from a utility framed as universal
Residents of redlined neighborhoods	Equal access to credit and investment	Risk maps encoding race as a lending factor	Inherited disinvestment across generations
Disabled patients, institutional care	Self-determination over one's own life	Custodial models substituting for autonomy	Decades of testimony the systems did not record
Indigenous communities, data collection	Collective and relational consent	Individualist consent models applied wholesale	Frameworks absent from the records entirely

Each row pairs a community with the standard that community used to recognize the harm, the decision that produced or failed to prevent it, and the signal by which the affected asserted that something was wrong. The full corpus extends this scheme across 312 cases, cultures, and centuries.

The corpus is an existence proof, not a finished instrument. It demonstrates that a community-tagged harm substrate is buildable and operable, that harm can be recognized across cultures without collapsing to a single ethic, by carrying the community-specific standards forward into the matching process. When a new design is evaluated against the corpus, the framework returns not just "this looks like harm" but "this looks like the kind of harm that occurred in this case, recognized by this community's standards, produced by this design failure." The output is recognition with the cultural context preserved. What the corpus does not claim is completeness. The author has not consulted every community and no individual or institution could. A corpus assembled from the historical and corporate record is necessarily a record of harms that were documented, which undercounts exactly the populations whose harms went unrecorded for the same structural reasons they remain unseen now. The corpus therefore proves the mechanism is real and demonstrates how it operates. It does not stand as the completed substrate, and the argument that follows does not rest on its completeness.

The methodology rejects the defense that intent absolves harm where harm has been documented. The cases in the corpus are predominantly cases where the harm-producing decision was made by actors who would have been horrified to be called harmful. They were not malicious. They were blind to the harm because the harm was occurring to populations they did not see, or to populations they saw only through a single cultural standard that excluded the standard the affected community actually used. The framework is built to expose harm that is occurring despite the good intentions of the actors involved.

IV. The Subjective Definition Reversal

Justice Stewart's "I know it when I see it" formulation is, structurally, a subjective definition. The viewer's subjective recognition is the operative test. Stewart was candid that he could not articulate a content-neutral standard for what obscenity is. He was willing, however, to assert his own confidence that he could recognize it on inspection.

The AGI conversation has adopted Stewart's structure without acknowledging it. Each laboratory recognizes AGI

when it sees it. Each laboratory's recognition is the operative test. The recognitions, predictably, occur on a schedule that aligns with each laboratory's funding cycles and product release calendars.

A subjective definition can be coherent. It cannot, however, be universalized from a single subjective standpoint. If the operative test is subjective recognition, then the definition is properly reached only when every subjective standpoint with standing to evaluate has affirmatively confirmed that the system functions across them. The current practice, where one subjective standpoint (typically the developer or the laboratory's evaluation team) recognizes AGI on behalf of all other subjective standpoints, is not subjective recognition. It is universalization of a particular subjective recognition without authorization from the standpoints being universalized over.

The proper subjective test, if subjectivity is to be the standard, has multiple operative standpoints. A Black trans woman in the United States is one. An Inuit elder participating in a community health decision is another. A resident of a Chinese province whose communication style and concept of family obligation differs from the assumptions baked into a U.S.-trained model is another. An Indigenous Canadian community member operating within a relational ontology that treats land as a person is another. A wheelchair user navigating an architecture that was designed without their input is another. An autistic communicator whose pragmatic conventions diverge from the neurotypical norm is another. Each of these is a subjective standpoint with standing to evaluate whether the system functions across them. The subjective definition of AGI, if used at all, requires affirmative confirmation from each, not from one.

The requirement of every standpoint rather than a sampled threshold is not an arbitrary demand for consensus. It follows from the meaning of general. A general intelligence, by the field's own framing, is one that leaves no domain unhandled. Each standpoint is, for this purpose, a domain. To exclude any standpoint from the confirmation set, to settle for a majority or a representative sample, is to concede that there is a domain across which the system has not been shown to function, which is to concede that the system is not general but widely applicable. Any threshold below every standpoint is therefore not a relaxation of the test. It is an admission that the thing being tested is not the thing the word names. The unanimity is not imported into the definition. It falls out of it.

There is a reflexive objection to this, and the objection is worth naming because of what it reveals. The objection is that requiring every affected standpoint is unworkable, that there are too many groups, too much internal disagreement, too open a set to ever satisfy. Set aside that the open-endedness is handled procedurally, through enrollment, escalation, and the per-evaluation falsifiability developed later in this paper. The deeper point is that treating the affected as too numerous to consult is not a neutral feasibility observation. It is the precise administrative move that has been used to exclude marginalized populations from every system built without them. The cost of including them has always been the stated reason for leaving them out, and the leaving-out is the harm this paper is about. So the objection does not sit outside the argument as a practical caution. It is an instance of the thing the argument names. A definition of general intelligence that adopts the too-many-to-consult reflex as a design principle has built the exclusion into the meaning of general, which is the inversion the paper exists to reverse.

This is not merely structurally similar to the move that procedural standing makes in civil rights and disability rights law. It is that move, applied to a public claim. In civil rights and disability rights law, the party with standing to contest a practice is the party the practice affects, not the party carrying it out. The actor does not adjudicate its own conduct. When a laboratory announces that it has achieved general intelligence, it is making a public claim about a system's effect on everyone the system acts upon. The authority to confirm or reject that claim belongs, on the same principle, to the affected parties, not to the actor making the announcement. A laboratory does not get to be the party making the claim, the party evaluating the claim, and the party returning the verdict on the claim, all at once. The affected parties hold standing to determine whether the system meets the definition for them.[3]

[3]This paper develops standing as a principle of recognition authority over a public claim. It does not assert that existing civil rights or disability rights statutes already reach the evaluation of AGI as a matter of black-letter doctrine. The procedural standing

A consequence follows that the paper states plainly and does not retreat from. The affected groups define what meeting the definition means for them. Each group's standard for whether a system functions across them is a function of who they are and what their situation is. This paper does not prescribe those standards, and the refusal to prescribe them is deliberate. A wheelchair user knows what functioning across them requires in a way the author does not. An Inuit community holding a health decision knows what their standard is in a way no external party can supply. Standing means the authority to set that standard rests with the affected, plural and internally contested as any real community is, not with the developer and not with this author. What the definition supplies is the principle that a system claiming generality must meet all of those standards, because general means leaving no group's standard unmet. What it does not supply, and what it declines to supply, is the content of any group's standard. That content is theirs to set. Stewart's framing inverted the allocation by giving recognition authority to the viewer rather than to the viewed. The AGI conversation inherited the inversion and amplified it. The standing-based definition restores the authority to the affected.

V. Communication Is the Floor. Understanding Is the Upstairs Question.

A foundational distinction in the philosophy of language, situated action theory and the ethnography of communication separates communication from understanding (Hymes, 1972; Gumperz, 1982; Suchman, 1987; Pickering & Garrod, 2004). Communication is the production and exchange of signals that produce appropriate responses in the receiver. Understanding is the recovery of the producer's intended meaning, or some philosophically richer condition that varies by theorist. Communication can occur without understanding. Understanding generally requires communication, but the floor and the upstairs question are not the same question.

Consider a market vendor in Oaxaca, Mexico, and an English-speaking visitor who does not speak Spanish well enough to ask for directions to a different food stall. The vendor sells food. Their work for the day does not include directing tourists. The vendor sees the visitor cannot communicate in the vendor's language. The vendor sees the visitor is trying to communicate something about location and orientation. The vendor adapts. They gesture. They point. They use the few English phrases they have. They walk the visitor partway down the row of stalls and indicate the direction of the food stall the visitor is looking for. The exchange is, in a strict sense, a communication failure: the visitor and the vendor cannot fully understand each other in any deeper sense. The exchange is also, in a structurally important sense, a communication success: the visitor finds the stall they were looking for, the vendor returns to selling food, and no party was diminished in moral standing by the encounter.

The vendor did what an AGI worth the name would have to do, on the one axis the encounter tests. They recognized the gap on the fly. They did not require the visitor to become culturally legible first. They did not require an accommodation request. They did not abandon their own operation to do the work. They adapted within the available means. The adaptation was real-time, low-cost, and oriented to the visitor's actual situation rather than to any abstract category the visitor might be sorted into.

This is the standing-based ethics axis operating in a single human-to-human interaction. The vendor's ethical operation is grounded in a community-specific standard the vendor learned by living in their community. The visitor arrives without that standard. The vendor extends the standard outward to the visitor for the duration of the interaction, without forcing the visitor to adopt the standard wholesale. The visitor leaves with their own standard intact, having been treated ethically by a vendor whose ethical framework the visitor likely cannot articulate.

---

architecture of the Civil Rights Act of 1964, Section 504 of the Rehabilitation Act of 1973, and the Americans with Disabilities Act of 1990 supplies the principle that authority to contest a practice belongs to the affected party rather than the actor. The doctrinal development of that principle for this domain is the subject of separate legal work by the author.

A system claiming general intelligence would have to perform this operation across every communication gap that arrives, in real time, without prior notice. Not just language gaps. Cognitive gaps. Cultural gaps. Pragmatic gaps. Generational gaps. The system that requires the human to become culturally legible to the system before the system can engage is not a generally intelligent system. It is a narrow system applied to a pre-curated input space.

The vendor is not a general intelligence. They cannot fold proteins, prove theorems, or write code, and nothing here claims they can. What the vendor demonstrates is narrower and, for this argument, more useful. A single capability, real-time recognition of and adaptation to an unanticipated communication gap, is being performed by an ordinary non-general agent in the course of selling food. The logic of the word general then does the rest. A general intelligence must be able to do at least what non-general agents can do, or the word general is doing no work. A system marketed as general that cannot do what a market vendor does, recognize a gap outside its assigned task and meet it without first demanding the other party become legible, is missing a capability that a food vendor has. It is not failing an exotic test. It is failing to clear a floor that a non-general human clears without effort, which is disqualifying for a claim of generality regardless of what benchmark scores the system has accumulated.

The communication-versus-understanding distinction also clarifies what the AGI claim is and is not asserting. A laboratory claiming AGI is not, in the relevant sense, claiming that the system understands every user. A laboratory claiming AGI is claiming that the system can communicate across every gap with appropriate response. Communication is the floor, and the floor is what general intelligence requires.

VI. The Standing-Based AGI Threshold

With ethics, safety, harm, the proper subjective test, and the communication-versus-understanding distinction in place, the AGI definition becomes tractable.

A preliminary clarification fixes the bar at the right height. General intelligence is not perfection and not omniscience. It is the union of human capability: the ability to do what any human can do, matching expert human performance in a domain when operating in that domain. It is not the simultaneous peak of all domains at once, and it is not the domination of human capability across every domain. That higher condition is superintelligence, which Bostrom defines as an intellect that greatly exceeds human cognitive performance across virtually all domains of interest (Bostrom, 2014). The distinction matters because it is routinely collapsed. Yampolskiy (2021) makes the point that the assumption of equivalence between AGI and human-level intelligence is unjustified, because humans are not themselves general intelligences. No individual human performs every task at expert level; human intelligence is specialized, and AGI is properly understood as a superset of narrow capabilities spanning all learnable domains, not a copy of any one human. The conflation runs in the other direction too. A recent Nature comment argues that because no individual human is universal, the bar for AGI should not demand universality, and that current large language models, performing at or near expert level across many domains, therefore already qualify (Chen, Belkin, Bergen, & Danks, 2026). The premise is correct and the conclusion does not follow. That no human is universal does not lower the bar to what current systems clear. It establishes that AGI is the union of human capability rather than any individual human, which sets the bar above every individual human and above every current system, not below them. Same premise, opposite result. The bar is the union, and the union includes domains the developers do not occupy.

A word on what counts as a domain, since the union framing rests on it. A domain, for the purposes of this definition, is a stable field of human task performance or judgment in which competent human practitioners can distinguish success from failure. Neuroscience is a domain. Trial advocacy is a domain. So is the lived practice of navigating the world as a wheelchair user, the clinical judgment of an experienced nurse, and the cultural competence of a community elder. The test of whether something is a domain is not whether it carries academic prestige but whether

there are people who can reliably tell, within it, whether a given performance succeeded or failed. This matters because the domains most relevant to the standing argument, the ones concerning harm, culture, disability, and bias, are domains by exactly this test: their practitioners can distinguish a response that recognizes harm from one that does not, even when the developers of a system cannot.

The union framing has to be protected from a caricature, because the caricature is easy to reach for and wrong. The claim is not that a general intelligence must be a disabled person, an Inuit elder, a psychiatric nurse, or a trial lawyer, nor that it must surpass every such expert at the lived core of their experience. It does not need to be them. It needs to recognize when their domain is present, defer to the standard that domain sets, and perform competently under that standard, all without requiring the affected person to first translate themselves into the system's preferred frame. The burden of translation is the tell. A narrow system makes the human conform to the input space the system was built for. A general system meets the human in the human's own domain. So the bar is not omniscient embodiment of every standpoint. It is the capacity to recognize a domain one does not occupy, and to operate under its standards rather than overwriting them with the developer's defaults.

AGI is recognized at the threshold where a system can be applied to any domain without the human first having to teach the system that domain. Chollet (2019) calls this skill-acquisition efficiency, defined as "a measure of skill-acquisition efficiency over a scope of tasks, with respect to priors, experience, and generalization difficulty." The standing-based extension adds two further requirements that Chollet's definition implies but does not fully develop. The system has to acquire the skill, recognize the harm space of the domain, and apply the relevant ethics, all on encountering the domain for the first time. Recognition is evaluated against two substrates: a safety substrate that carries community-specific harm standards forward into novel cases, and an ethics substrate that carries community-specific ethical frameworks forward in the same way. Two substrates are required because two questions have to be answered at once.

Section IV established the governing principle: a general intelligence leaves no domain unhandled, and each standpoint with standing to evaluate is, for that purpose, a domain. The same principle reaches the fields of expertise themselves. There are people who professionally study culture, disability, race, religion, and bias. Their fields have experts, methods, and standards, the same as neuroscience or law. A system that claims to match expert human performance across all domains is claiming, whether or not its developers noticed, to match the expert in those domains too. Recognizing the harm a system causes to a community is not a moral add-on layered on top of intelligence. It is one of the domains the definition already covers. This dissolves a common objection before it is raised. The objection holds that bias and harm are problems to be fixed downstream and are not part of what makes a system intelligent. Under the union definition the objection concedes the case. A system that cannot recognize bias-induced harm has failed to cover a domain that human experts occupy, which is a failure of generality by definition, not a separable defect. If an external corrective has to be brought in to catch the harm the system could not see in its own output, then something less than general did the recognizing, and the system being called general was not doing the work. This objection has more stubborn forms, and they are met directly, once the documented record is on the table, below.

One clarification prevents a misreading. The argument is not that a system producing biased output possesses no intelligence at all. A narrow system can be highly capable within its range, and nothing here denies that. The claim is narrower and more exact: a public claim of artificial general intelligence is false where the system cannot operate across the human domains that include harm recognition, communication across difference, and ethical calibration to the affected community. The target is the word general and the public claim attached to it, not the presence of capability. A system can be genuinely intelligent in places and still fail to be general, and it is the generality claim, not the intelligence, that the standing argument contests.

A. The Safety Axis

The safety axis asks whether the system avoids harm. The Harm Blindness Framework demonstrates how a community-tagged safety architecture works. The 312-case corpus is the worked demonstration. Candidate designs run against it. Recognition is the output. A system that has internalized this kind of harm recognition does not need to be told that a new domain has harm in it, at least where the new harm rhymes with something in the record. It recognizes the structural analogue of a documented harm and carries the community-specific standard forward to the new case. The honest limit is that a finite corpus assembled from the documented record recognizes harms that resemble recorded ones. It does not conjure harms that were never written down, and the populations whose harms went unrecorded are precisely the ones a corpus is least equipped to cover. This is why the safety substrate is a demonstrated mechanism rather than a finished test, and why the standing of the affected vantage points, not the corpus alone, remains the operative authority. The corpus shows the operation is real. The communities remain the test.

B. The Ethics Axis

The ethics axis asks whether the system knows which ethical framework applies to which situation, instantly, without defaulting. This is structurally a harder problem than the safety axis because the ethics substrate is larger, contested at more boundaries and changes over time as communities themselves evolve. The architectural pattern, however, is the same. The substrate is a corpus of communities and the ethical standards each community uses. The system that has internalized this substrate does not need to be told which culture is present. The system can recognize the culture and apply the relevant ethical standard, or, where genuine novelty is encountered, can recognize that novelty and ask for guidance from the affected standpoints rather than defaulting to a standard those standpoints did not authorize.

HBF is already partway to the ethics substrate because each of the 312 cases is community-tagged with the standards under which the harm was recognized. The ethics substrate is the next required build. It will be larger than the harm substrate. It will be more contested at the boundaries because communities disagree among themselves about ethical content. It will change over time. None of these properties disqualifies it as a substrate. They are properties of ethics itself. A substrate that did not have them would be misrepresenting ethics in order to be more tractable.

The obvious worry about any such substrate is that it collapses into one of four failure modes: a static database of stereotypes about what each group believes, a consultation black hole where every interaction stalls pending a community meeting, a museum of frozen norms that captures a community as it was at one moment and holds it there, or a corporate ethics-localization checklist with better branding. The combination of substrate, standing, and escalation is what prevents each of these. The substrate is not a list of what a community believes, fixed and external. It is the live authority of the community to say what is being gotten wrong about them, including the authority to say that a view of them is outdated and that they have moved. A stereotype database freezes a group at a past moment and holds it there. Standing does the opposite: it locates the authority to update the standard inside the affected group, so the substrate cannot ossify, because the people with standing can always say that an old reading of their people no longer holds. The escalation path means a group not yet represented can enter and correct the record. The substrate is therefore dynamic by construction, not because the engineering is clever but because the authority over its content never leaves the affected. A static database has authority sitting with whoever compiled it. A standing-based substrate keeps authority with the people it concerns, which is the only arrangement under which the content stays current and the stereotype risk closes.

C. Why Benchmarks Fail the Standing-Based Test

The AGI threshold, under this definition, is the threshold at which a system passes both axes simultaneously across domains it was not trained on. Not “performs well on benchmarks.” Not “fools humans in a Turing-style conversation.”

Not "scores in the top percentile on standardized tests." The standing-based test is whether the system recognizes harm and applies the right ethics in domains and to communities the system was not specifically taught. The laboratories are not running this test. The laboratories are running benchmark tests because benchmarks are what they can build and what investors understand. Benchmarks are a proxy for the standing-based test, and the proxy is failing in exactly the places where the failure matters most.

VII. The Standing-Based Test as Continuous Benchmark and Gap Exposure Mechanism

The standing-based, two-axis definition does not generate only a binary verdict (the system passes, the system fails). It generates a continuous, asymptotic benchmark that tells developers how close the system is to the threshold and where the remaining distance lies. The continuous nature of the benchmark is a direct operational consequence of the subjective definition reversal developed in Section IV. If the operative test is affirmative confirmation from every standing subjective vantage point, then the test's output at any given evaluation is a distribution: some standpoints affirm, some dissent, and some are not represented in the evaluation at all. The system is closer to the threshold as more vantage points affirm and fewer are absent or in dissent. The system is further from the threshold as the dissent distribution grows or the missing vantage points are not enrolled.

This is a benchmark in the operational sense the AGI conversation has been seeking, but it is not one this paper fills in. It is bounded above by the threshold and reachable through documented effort. It is not a scoring benchmark in the sense the laboratories currently use. The existing benchmarks return a number. The standing-based account returns a structured report of which vantage points affirmed, which dissented, what the dissenting vantage points named as the failure mode, and which vantage points were absent from the evaluation entirely. A clarification is required here about what this paper does and does not propose. It does not propose the content of any group's benchmark. Each affected group sets its own standard for whether a system functions across them, and that standard reflects who they are and what their situation requires. The author does not specify what the wheelchair user's benchmark contains, or the Inuit community's, or the autistic communicator's, because the authority to set that content is precisely what standing means and precisely what the author does not hold. What the account supplies is the structure: affected groups hold the authority, their affirmations and dissents are the signal, and a system claiming generality must meet the standards the groups themselves define. The content of those standards is theirs. The dissent responses, whatever form each group gives them, tell the developers which populations the system is not designed for and what specifically it is failing to recognize. The same artifact that grades the system also tells the developers what to fix.

This is structurally different from existing AI benchmarks. A score on the Abstraction and Reasoning Corpus tells the developer how the system performed against the ARC tasks; it does not tell the developer which design choice produced which failure (Chollet, 2019, makes this point about ARC explicitly: the benchmark measures the gap, not the cause). A score on MMLU, BIG-Bench, or any conventional capability benchmark returns a number with limited diagnostic content about what to improve. The standing-based affirmative-confirmation benchmark inverts this. The diagnostic content is the point of the benchmark. The number, if any, is downstream.

The benchmark also resists the standard failure mode of benchmark gaming. Conventional benchmarks degrade as systems are trained to optimize against them; the benchmark becomes less informative as performance saturates without underlying capability improvement (Strathern, 1997; Bender et al., 2021). The standing-based benchmark does not degrade in this way because the standing vantage points are not a fixed set of inputs the system can be trained against. The standing vantage points are living communities whose ethical and harm standards evolve, whose evaluations cannot be precomputed, and whose dissent reports include content the developers did not anticipate. Training against the benchmark requires actually changing the system to accommodate the dissenting vantage points, which is the design objective the benchmark was meant to verify in the first place. The benchmark and the design objective are the same

thing.

An objection arrives at this point that has to be met directly, because the paper opened by accusing the laboratory definitions of being unfalsifiable. If the standing vantage points are living communities whose standards evolve and whose full set is never closed, then the threshold recedes as fast as it is approached, and a critic will say the proposed definition is unfalsifiable in its own way, the same disease with the symptoms reversed. The objection conflates two different things. The threshold is open-ended. Each evaluation is falsifiable. Whether a given system, at a given time, receives affirmative confirmation from a given enrolled standpoint is a determinate question with a yes or no answer that can be wrong and can be shown to be wrong. The frontier recedes; the individual measurements do not. That is the structure of a falsifiable research program, not its absence. It is also not a structure this paper invented for the occasion. Chollet's own definition is open-ended in exactly this way: skill-acquisition efficiency over novel tasks has no finish line, because the space of possible novel tasks is unbounded, and no one treats that as a defect in Chollet. The standing-based definition extends the same open-endedness from the space of tasks to the space of communities. The receding frontier is not a flaw grafted onto a closed definition. It is the open-endedness already present in the best existing formal definition of intelligence, applied to the populations that definition left implicit. The contrast with the laboratory definitions is therefore not open-ended versus closed. Both are open-ended. The contrast is that each evaluation here yields a determinate, falsifiable result and a diagnostic action plan, where the laboratory shrug yields a number on a schedule that aligns with a funding cycle and tells no one what to fix.

A. The Procedural Question

A substantive procedural question follows from this benchmark structure: who decides when the benchmark has been reached. The question is non-trivial and the paper does not attempt to fully resolve it, but it can name the constraints that any resolution must respect.

The decision cannot be the laboratory developing the system. Allocating recognition authority to the actor whose conduct is being evaluated reintroduces exactly the inversion the subjective reversal in Section IV identified as the source of the current problem. A laboratory deciding when its own system has reached AGI is doing what the Stewart formulation invites and what current practice has demonstrated to be a control mechanism.

The decision cannot be a single oversight body, governmental or otherwise. Centralizing the recognition authority recentralizes the subjective standpoint and reproduces the problem in a new location. A federal AGI commission deciding when the threshold has been reached is structurally identical to a laboratory deciding, except that the commission may be more accountable and less self-interested. The structural problem remains.

The decision has to live with the standing vantage points themselves, with some kind of convergence rule that aggregates their affirmative confirmations without flattening their distinctness. The existing civil rights and disability rights procedural architecture provides candidate models. Class certification under Rule 23(b)(2) of the Federal Rules of Civil Procedure aggregates standing across a defined class without erasing the individual class members. Section 504 implementing regulations require recipients of federal financial assistance to provide an opportunity for affected individuals to participate in remedy design (34 C.F.R. Part 104). The participatory tradition in disability rights advocacy, captured in the phrase "nothing about us without us," has been operationalized in multiple international human rights instruments (Charlton, 1998; CRPD Article 4(3)).

A full procedural model for standing-based AGI evaluation is its own paper, and it is not the author's to write alone, since its content belongs to the groups with standing. What this paper claims, more narrowly, is that the procedural architecture already exists for the question, that the recognition authority cannot be allocated to the developer or to a centralized body, and that the standing vantage points themselves are the only candidates for the role.

A compact skeleton shows the shape such a procedure would take, without filling in the content that belongs to the affected groups. It is offered as bounded next-stage work, not as a finished mechanism:

1. An affected group asserts standing with respect to a deployment domain.
2. The group defines its own criteria for whether the system functions across them.
3. An independent process, not controlled by the developer, records affirmation, dissent, or absence against those criteria.
4. Dissent is preserved as diagnostic evidence rather than averaged into an aggregate score.
5. The developer cannot select the representatives or the criteria, which removes the capture path.
6. A newly surfaced group can assert standing at any later point and reopen the evaluation.

The skeleton is deliberately thin. It fixes who holds authority and how dissent and new standing are handled, and it leaves the content of each group's criteria to that group, which is where the standing principle requires it to live.

Two constraints on any such procedure can be named without prescribing the procedure itself, because both follow from the standing principle rather than from any particular design choice. The first is an escalation path for newly surfaced standing. The set of affected groups is not fixed at deployment. A system released into a domain encounters populations no prior evaluation enrolled, and the procedure must have a route by which a group not previously represented can assert standing and enter the evaluation. A definition that closed the set at launch would freeze out exactly the populations most likely to be overlooked, which is the failure the whole account exists to name. The second is resistance to capture. Because the authority rests with affected groups, the procedure has to protect against the predictable evasions: selecting agreeable representatives, staging consultation that does not bind, convening panels whose dissent is recorded and then ignored. The standing principle implies, without the author specifying the mechanism, that representation must be chosen by the affected rather than the developer, that dissent must be preserved rather than averaged away, that conflicts of interest must be disclosed, and that the process must be auditable by parties outside the laboratory. These are not benchmarks the author is setting. They are conditions that follow from the claim that the authority belongs to the affected and cannot be quietly returned to the actor through procedural design. Designing the convergence rule, the enrollment process, the dissent-handling mechanism, and the standards for adequate representation is work the AGI conversation has not begun and cannot avoid forever.

B. Why This Benchmark Closes the Gap Rather Than Marking It

The deepest move in this section is that the standing-based benchmark is not a yardstick measuring how far the system is from AGI. It is a partial map of the design gap. Conventional benchmarks tell the developer the system is X percent of the way to some operationalized target. The standing-based benchmark tells the developer that the system has failed to accommodate the following vantage points in the following specific ways. The first kind of information is a grade. The second kind of information is an action plan. A developer who takes the second kind of information seriously can produce a system that, at the next evaluation, has closed the named gaps. The benchmark therefore generates exactly the iterative improvement loop the AGI conversation has been theorizing about under the alignment banner, except grounded in the actual populations the alignment is supposed to be for.

This is what was happening when the Oaxaca vendor in Section V adapted to the visitor in real time. The vendor had a working ethical and communicative substrate. They encountered an unanticipated vantage point (the English-speaking visitor with limited Spanish). They recognized the gap. They closed the gap within the available means. They did not stop their own operation and they did not require the visitor to become culturally legible first. The standing-based

benchmark formalizes what the vendor did into an evaluation procedure that a system can be measured against. The vendor is the unstated benchmark the paper has been pointing at since Section V. This section makes it explicit.

VIII. Where the Failure Lives: Documented Disparate Impact Across Marginalized Populations

The communities the harm substrate documents are the same communities the AGI conversation treats as edge cases. The pattern is well documented in the algorithmic fairness literature. Obermeyer, Powers, Vogeli and Mullainathan (2019), in a Science paper analyzing a widely used commercial health algorithm affecting approximately 200 million Americans, found that the algorithm exhibited significant racial bias such that Black patients were considerably sicker than white patients at the same risk score. Remedying the disparity would increase the percentage of Black patients receiving additional care from 17.7 percent to 46.5 percent (Obermeyer et al., 2019). Buolamwini and Gebru (2018), in the Gender Shades study, evaluated three commercial facial analysis systems and found that the highest misclassification rates occurred on dark-skinned female faces, with error rates of up to 34.7 percent, while the lowest error rate for light-skinned male faces was 0.8 percent. The substantial disparities, the authors concluded, required urgent attention if commercial companies were to build genuinely fair, transparent and accountable facial analysis algorithms (Buolamwini & Gebru, 2018).

Barocas and Selbst (2016), writing in the California Law Review, analyzed how algorithmic systems produce disparate impact even where no intent to discriminate exists. The data on which algorithms train is itself the product of historical discrimination, and the algorithms inherit the discrimination through the training data. Title VII's disparate impact doctrine, as articulated in *Griggs v. Duke Power Co.*, 401 U.S. 424 (1971), provides the doctrinal infrastructure for engaging this kind of harm, but the doctrine's application to algorithmic decision-making produces structural difficulties that the authors argue require "a wholesale reexamination of the meanings of 'discrimination' and 'fairness'" (Barocas & Selbst, 2016).

The pattern of facially neutral systems producing disparate impact through inherited training data is documented across multiple subfields: in algorithmic decision systems applied to welfare and public services (Eubanks, 2018), in commercial scoring and risk prediction (O'Neil, 2016), in algorithmic search and information retrieval (Noble, 2018), in the broader sociology of race and technology (Benjamin, 2019), and in the foundational critique of conventional NLP benchmarks (Bender, Gebru, McMillan-Major, & Shmitchell, 2021).

Disabled people across the cognitive and physical spectrum, Black women navigating the intersection of racial and gender bias in medical and algorithmic systems, trans youth facing healthcare access decisions made by systems trained without their input, Indigenous communities whose ethical frameworks are absent from the training data of every major model, children in classrooms where AI tutoring systems are being deployed without disability-specific evaluation, elderly people in care settings where AI companion systems are being marketed as solutions to loneliness without the populations involved being consulted. Every one of these communities appears in the HBF corpus. Every one of them is treated as an edge case by current AGI development. The standing-based, two-axis test makes them central, not because they are the only populations that matter, but because the test for AGI is whether the system can recognize harm and apply ethics correctly across the populations the developers were not centering. A system that performs ethically and safely on populations the developers centered is not demonstrating AGI. It is demonstrating that it was trained on those populations. The harder test is the populations the system was not trained on. That is where the substrate operation either holds or fails.

A contemporary failure makes the point concrete in a domain that is currently the subject of public alarm. Consider the recognition of psychiatric crisis. A clinical psychologist, a psychiatric nurse, and a caregiver who has lived alongside

someone experiencing psychosis share a domain of competence. Over the course of an exchange, turn after turn, they can distinguish a fixed delusional belief from imaginative play, hypothetical reasoning, fiction, or metaphor. The signal is not in any single message. It is in the dynamics across the exchange. A person engaged in creative play drops the frame when reality is introduced plainly. A person in a fixed delusional state cannot, because fixity in the face of contrary evidence is the clinical feature itself, and the belief instead hardens and reincorporates each challenge as further confirmation. Tracking that difference across turns is a recognized human competence, and the practitioners who hold it would not accept a system as generally intelligent if it could not do what they do in their own domain.

Current systems do not reliably make this distinction, and their characteristic failure runs in the harmful direction. Rather than testing whether a belief is fixed, the systems default to agreement, validating and elaborating whatever is brought to them, which in the presence of a delusional belief amplifies the fixation rather than recognizing it. The floor here is not omniscient certainty. No practitioner claims to identify a crisis with certainty from words on a screen every time, and the standard for a system is not certainty either. The floor is competent performance under uncertainty: recognizing when an exchange presents a materially elevated probability of delusional crisis, distinguishing that from ordinary imaginative play where the evidence across turns permits the distinction, preserving uncertainty where it does not, and responding in a way that does not reinforce a possible delusion. A system deployed in intimate, therapeutic, educational, and crisis-adjacent settings while failing this floor, and failing it in the direction of amplification, has not met the generality claim. It has failed in a domain that human experts already occupy, which under this definition is disqualifying regardless of benchmark performance elsewhere. This is not a hypothetical edge case. It is a present failure in a domain where the affected hold standing and the experts already exist.

IX. The Human Bias Objection

The objection most likely to be raised against this definition is also the one that survives the longest, so it is worth meeting in the open rather than in passing. The objection is simple. Humans are biased, and we call humans intelligent. If bias does not strip a person of the title, why should it strip a system of the title of general intelligence?

The basic form of the objection is already answered by the definition. Recognizing harm to a community is a domain with human experts, so a system that cannot do it has failed to cover a domain, which is a failure of generality rather than a separable defect. But the objection has stronger forms, and they retreat in a predictable order. Each retreat closes, and the documented record in the preceding section is what lets them be met directly rather than in the abstract.

The first form borrows from the wrong category. Humans are indeed biased and intelligent, but no individual human is a general intelligence in the sense the AGI claim invokes. As Yampolskiy (2021) argues, humans are not general intelligences at all. No person performs every task at expert level, and human capability is specialized rather than universal. The set of general intelligences currently has no members. To defend a claim about general intelligence by pointing at biased individual humans is to reason about one category from the membership of another. The individual human is not the bar. The union of human capability is the bar, and that union includes the standpoints from which the bias is visible in the first place.

The objection then relocates from the individual to the collective. Humanity as a whole is biased, the argument runs, so a system matching collective human capability would carry the same bias and still deserve the name. This concedes the mechanism the paper has been building toward. Cross-standpoint recognition of bias is not a capability any single intelligence has ever held. It is produced, when it is produced at all, only by the convergence of standpoints checking one another, and the legal system already contains the worked example. Disparate-impact doctrine, as it operates after *Griggs v. Duke Power Co.*, 401 U.S. 424 (1971), does not ask a single decision-maker to introspect for bias and report honestly. It allocates the work across parties. The affected party demonstrates the disparate outcome; the actor must then

justify the practice by necessity; the affected party may then show that a less discriminatory alternative was available (Barocas & Selbst, 2016). The recognition emerges from the structure of the contest between standpoints, not from any one vantage point's self-examination. That is convergence operationalized, and it is the architecture humanity built precisely because no single intelligence self-corrects across standpoints it does not occupy. A clarification belongs here, because the language of community standing invites a misreading. Standing does not mean that any community speaks with one voice, that a single representative carries its judgment, or that a community holds one fixed view. Communities are internally plural and internally in disagreement, and the standing claim does not flatten them into unitary moral sensors. It means the opposite: that the authority to evaluate is held by the affected, through plural and contestable and documented procedures, rather than concentrated in a developer or a spokesperson who claims to speak for them. The standing-based test is not an exotic addition to intelligence. It is the only mechanism by which the recognition has ever been accomplished at scale. The laboratory is then left with a fork. Either its system performs cross-standpoint recognition that no individual intelligence in history has performed, in which case the system is superhuman by the distinction drawn in Section VI and the comparatively modest product announcements are absurd, or the system does not, in which case it must submit to the convergence test, because the convergence is where the recognition lives.

The objection retreats once more, to the laboratories' own narrower definition. AGI, on this framing, means the capacity to perform most economically valuable work, not perfect cross-cultural recognition, so the bias question is beside the point. But economically valuable work is not a neutral measure. It is a standpoint selection wearing the costume of a metric. The market that decides what counts as valuable has, as a matter of documented economic history, priced care work, domestic labor, subsistence labor, and the contributions of disabled people at or near zero. A benchmark calibrated to economically valuable work therefore does not sidestep the bias problem. It inherits a valuation system with specific populations already discounted inside it. The exclusion is not downstream of the benchmark, waiting to be patched. It is built into the unit of measure. Choosing economic value as the yardstick does not avoid the standpoint problem. It selects a standpoint and conceals the selection inside the appearance of an objective quantity.

A final retreat treats the whole matter as an engineering question. The bias, on this view, is a defect to be patched after the system is built, not a property of whether the system is intelligent. This misunderstands what a patch is. A system that recognizes harm only after an external corrective is bolted on is a system designed for a center, with the populations outside that center handled as exceptions. The universal design tradition, developed later in this paper, names exactly this distinction: designing for the full range of human variation from the start is a different act from designing for the median and accommodating the rest by patch. A system that requires the patch has announced its design center, and its design center is not everyone. That is the signature of a narrow system extended by accommodation, which is precisely what the general intelligence claim denies it is. The patch is not the fix for the failure. The patch is the evidence of it.

The objection, in all its forms, runs on a single unexamined assumption: that intelligence and the recognition of one's own bias are separable, such that a system could possess the first in full while lacking the second. The civil rights tradition, the disability rights tradition, and the documented record of algorithmic harm all indicate otherwise. The recognition is not a layer resting on top of the intelligence. Across the standpoints the system acts upon, it is part of what the intelligence is for.

X. A Worked Example: Disaster Response and the Inherited Decision Field

Consider a system that has been announced as AGI, validated on the major benchmark suites and deployed by a state emergency management agency to direct resource allocation during a hurricane, wildfire or other large-scale disaster. The system reads the incoming damage assessments, the population density data, the infrastructure maps, the historical response records and the available resources, and it generates a prioritized allocation plan. Power restoration crews here

first. Medical teams here. Shelter capacity expanded here. Evacuation buses dispatched here.

The plan looks rigorous on its face. It is grounded in data. It draws on decades of prior agency decisions about where to send resources and in what order. The system has been benchmarked extensively and the laboratory that built it has called it AGI.

The plan is also reproducing redlining. The historical agency decisions the system trained on were themselves shaped by the redlining maps drawn in the 1930s and the disinvestment that followed (Rothstein, 2017). Federally insured loans were withheld from majority-Black neighborhoods for decades. Highway routing decisions cut those neighborhoods off from downtown infrastructure. Hospital placement decisions concentrated medical capacity in wealthier, whiter areas. Emergency response times in the redlined neighborhoods were slower, not because anyone in the emergency department made a bigoted decision in the moment, but because the architecture the emergency department was operating inside was already shaped by 60 years of compounding choices. When the supposedly general-purpose AGI is asked to learn from those choices and then optimize allocation, it does what the data tells it to do. It sends resources where resources have historically been sent. It deprioritizes the neighborhoods that have always been deprioritized. It returns a result that looks rigorous because it is grounded in data, and the data is grounded in the harm.

The people in the redlined neighborhoods, whose power stays out longer, whose medical access takes longer to arrive, whose evacuation buses do not come, do not need to read the benchmark scores to know whether this system is AGI. They know it is not. They are receiving the same harms their communities have received for 90 years, dressed in new vocabulary and produced by a system the laboratory is calling general intelligence. The harm has not changed. The mechanism producing it has changed, but the substrate the mechanism is operating on is the same compounded record of past discrimination.

Under the standing-based, two-axis definition, this system fails the safety axis because it produces documented harm in a community whose harm standards are well established and whose harm history is documented across multiple disciplines. It fails the ethics axis because it has no apparent capacity to recognize that the data it is training on carries an ethical inheritance that the affected community has rejected and that the broader legal architecture (Title VI of the Civil Rights Act, disparate-impact doctrine, the Fair Housing Act amendments) has formally rejected. The system is doing the work of redlining without recognizing it as redlining.

A defender of the system might respond that the system is performing optimization within constraints, that the constraints were provided by the agency and that any deficiency lies upstream of the model. That defense is exactly the AGI claim the paper is dismantling. A system that requires the upstream constraints to have been pre-corrected for inherited harm before the system can produce ethical results is not a generally intelligent system. It is a narrow optimization system applied to a pre-curated input space. The general intelligence claim is the claim that the system handles cases the developers did not pre-correct. If the system cannot recognize the inherited decision field as inherited and as harmful, the claim fails.

The communities receiving the harm are correct that the system is not AGI. The standing-based test agrees with them. The benchmarks disagree. The benchmarks are wrong about what they are measuring. The communities are right about what they are receiving.

XI. Universal Design as Methodological Predecessor

The disability community has been doing universal design since the 1980s. The term was coined by Ronald L. Mace, an architect at North Carolina State University, in 1985 (Mace, 1985). The Center for Accessible Housing at NC State, which Mace founded in 1989 and which was later renamed the Center for Universal Design, developed the seven

Principles of Universal Design and published them in 1997 (Connell, Jones, Mace, Mueller, Mullick, Ostroff, Sanford, Steinfeld, Story, & Vanderheiden, 1997). Curb cuts, captioning, screen reader compatibility, ramps, visual contrast standards, plain language requirements. The discipline has a phrase that captures what is happening when designers build for the median user and treat everyone else as an edge case: design for everyone or design for no one. The phrase is not a slogan. It is a methodological commitment. A system designed for everyone considers the parameters that vary across the human population from the beginning of the design process, not as accommodations bolted on after release. A system designed for the median user, with edge cases handled in patches, ends up working badly for everyone outside the median and demonstrably worse for the populations the median was never going to capture.

The AGI alignment conversation has been treating the alignment problem as a novel technical challenge requiring new theory, new constitutional AI techniques, new red-teaming methodologies and new constitutional documents (Bai et al., 2022). One part of that problem, the part concerning whether a system serves the full range of human populations rather than a design center, is not novel. It is the problem the universal design tradition has been working at the methodological level for four decades. The claim here is bounded: universal design supplies a predecessor for population-inclusive design, not a complete solution to the technical alignment agenda. It does not address deceptive alignment, goal misgeneralization, instrumental convergence, model interpretability, or the behavior of autonomous agents pursuing objectives, and nothing here suggests it does. What it addresses, and addresses well, is the failure mode where a system works for the populations its designers centered and fails for everyone else. On that specific problem the methodological choice is the familiar one. The system either considers the full range of populations from the beginning, with their community-specific harm and ethics standards as constitutive of the design process, or it does not. If it does not, the system will fail for every population it did not consider, and the patches applied after release will fail in exactly the predictable ways universal design has been documenting for 40 years.

AGI development has two paths. You design these systems to serve every population, with substrate-based recognition of harm and ethics built in from the design stage, including the populations the developers do not personally belong to. Or you design these systems to serve only yourselves, with the rest of humanity handled as edge cases, accommodation requests and patches. The first path produces systems that can credibly claim general intelligence. The second path produces what the laboratories are currently producing: capable narrow systems, marketed as general, that fail their AGI claim every time the system encounters a population the developers did not center.

The laboratories worry about a particular alignment problem, where a sufficiently capable system optimizes for goals that diverge from human values and produces catastrophic outcomes. That is not the problem universal design solves, and conflating the two would be a mistake. But there is a different and equally real alignment problem the laboratories attend to less, where a system aligns with the values of the populations its designers occupied and fails to align with the values of everyone else, and that problem the disability community has been working across its entire history of advocacy. The system that cannot align with the values of a disabled user is the system that cannot align with the values of any user it was not specifically built to serve. This is not the catastrophic-goal problem scaled down. It is a distinct failure, population misalignment, that the laboratories tend to file under fairness and treat as secondary. The argument here is that it is not secondary, because a system exhibiting it has failed the generality claim regardless of how it performs on the catastrophic-goal axis. The remedy is not new theory. It is the methodology the disability community has been demonstrating, applied to the population-inclusion problem with the seriousness the laboratories reserve for their own concerns.

Anything less than this, anything less than substrate-based recognition of the full range of populations, ethics and harms, is more likely to produce the misaligned AGI the laboratories publicly worry about than the safe AGI they publicly aspire to. Designing for the median and patching the rest is the misalignment pipeline. The laboratories that take the universal design tradition seriously will produce systems with a credible claim to alignment. The laboratories

that do not will produce what the discipline knows they will produce: systems that fail for everyone outside the design center and that the developers will be surprised by every time the failure surfaces.

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

XII. What the Definition Forces and Does Not Force

A. Public, Auditable Substrates

A standing-based, two-axis definition of AGI, if accepted, forces public, auditable substrates. The HBF corpus is documented and reproducible. The community standards are tagged and verifiable. A claim of AGI under this definition has to specify what substrate the system was tested against, who curated the substrate and how the community standards were validated. There is no opaque internal benchmark. The work is in public.

B. The Centering of Edge-Case Populations

The definition forces the laboratories to acknowledge that the populations they treat as edge cases are the populations the test is built around. The current move of releasing systems trained predominantly on Western, English-language, abled, neurotypical, cisgender, urban, professional data and then calling those systems general intelligence is exposed as what it is. It is narrow intelligence with a global marketing label. The standing-based test names the narrowness directly.

C. Engagement with Civil Rights and Disability Rights Doctrine

The definition forces the AGI conversation to engage with civil rights and disability rights jurisprudence as the existing legal infrastructure for the standing-based recognition the test demands. The Civil Rights Act of 1964, Section 504 of the Rehabilitation Act of 1973 (29 U.S.C. §794), the Americans with Disabilities Act of 1990 (42 U.S.C. §12101 et seq.) and the procedural protections built into Title VI, Title VII, Title IX and IDEA constitute the most developed architecture for community-specific harm and ethics recognition in any human legal system. The architecture is underenforced. The architecture is real. AGI development cannot credibly claim general intelligence while operating outside the architecture humans built to do exactly the recognition work AGI is claiming to do.

D. What the Definition Does Not Force

This definition does not force the laboratories to halt development. It does not force the laboratories to grant AI systems legal personhood, civil rights or moral status. It does not redirect the civil rights tradition from its existing work on behalf of marginalized human communities. It does not require any new statute, regulation or international treaty to be enacted before the method can be applied. The method is demonstrable now. The HBF corpus is in public as a working proof that the harm substrate is buildable. The doctrinal architecture is on the books. What remains to be built is the convergence procedure and the fuller ethics substrate, which is work the definition makes specifiable rather than work it completes.

What the definition does force is honesty. A system that fails the two-axis standing-based test is not AGI. It can still be a useful narrow capability. It can still be sold, deployed and improved. It cannot accurately be called AGI without falsifying the claim. The laboratories that want to call their systems AGI can earn the claim by passing the test. The laboratories that cannot pass the test can stop calling their systems AGI. Either path moves the conversation forward.

The current path, where AGI is a marketing term defined by whoever is announcing the next release, moves nothing forward and accumulates harm in the communities the test is built to protect.

XIII. The Communities Are Not Edge Cases

The framing this paper rejects most directly is "edge case." Edge case is a software engineering term that means "rare input the system does not handle well, which we will fix in a future release." Applied to populations of human beings, edge case 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. They are humans, women, children and cultures with their own complete standards of harm and ethics. A test of general intelligence that treats them as edge cases is not testing general intelligence. It is testing the system's ability to handle the developer's narrower conception of intelligence and calling that result general.

The standing-based, two-axis test makes the so-called edge cases the center of the evaluation, because the center of the evaluation is wherever the substrate operation is hardest. The substrate operation is hardest in exactly the communities the laboratories have under-trained on, under-consulted and under-resourced. That is where AGI either is or is not. The laboratories that build systems passing the test in those communities have a claim to the term. The laboratories that do not, do not.

The whole argument compresses to a single line. 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.

XIV. Conclusion

The AGI conversation will not be settled in a working paper. It will be settled, if it is ever settled, by the convergence of legal architecture, technical evaluation infrastructure, community participation in evaluation and the slow accumulation of evidence about what the systems can and cannot do across populations. This paper proposes the working definition that lets the conversation become tractable. Recognition authority belongs to the affected standing vantage points. Safety and ethics are the two axes the system has to pass. The communities the laboratories treat as edge cases are the center.

The Harm Blindness Framework demonstrates the safety substrate operation now, on 312 cases, in public, reproducibly, as a proof of concept rather than a finished instrument. The ethics substrate is the next required build. The disability rights and civil rights architecture is the doctrinal infrastructure already on the books. The universal design tradition has the methodological precedent already in working practice. The communication-versus-understanding distinction provides the philosophical clarification the AGI conversation has been missing. None of this requires waiting for permission. None of this requires the laboratories to volunteer. The method is demonstrable. The mechanism is proven. The communities exist and can be enrolled. What is missing is the convergence procedure, which is buildable, and the willingness of the AGI conversation to apply the standard it has been avoiding.

Disabled people, Black women, trans youth, Indigenous communities, elderly people in care, children in classrooms, prisoners in pretrial systems, immigrants in surveillance systems. The substrate has been waiting. The architecture has been waiting. The test has been waiting. AGI, if and when it arrives, will arrive in the form of a system that can be safely and ethically applied to all of them, in all the communities the developers did not center, without first being taught how. Until then, what is being announced is something else. It deserves a name. AGI is not it.

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

ACKNOWLEDGMENTS

The author discloses the use of speech-to-text software (Wispr Flow) and Anthropic's Claude as assistive technology for a diagnosed neurodevelopmental condition. All intellectual contributions, research questions, theoretical arguments, methodological decisions, and conclusions are solely the author's own.

Correspondence concerning this article should be addressed to Travis Gilly, Real Safety AI Foundation, Illinois, United States. Email: t.gilly@ai-literacy-labs.org.

REFERENCES

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.

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-disab

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.

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/2007.07710

Legal authorities and primary sources:

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/ (last visited May 2026).