Viewpoint

What Does Your AI Mean by “Flourishing”? The Case for Disclosing and Benchmarking the Values in AI Alignment

Alexandre Castonguay, PhD

Faculty of Nursing, Université de Montréal, Montréal, QC, Canada

Corresponding Author:
Alexandre Castonguay, PhD
Faculty of Nursing
Université de Montréal
2375, chemin de la Côte-Sainte-Catherine
Montréal, QC, H3T 1A8
Canada
Phone: 1 514 343 6437
Email: alexandre.castonguay.2@umontreal.ca

Abstract

Who could be against human flourishing? Every AI developer promises that its systems will serve humanity, and governments promise the same. It sounds like consensus. It is not: behind the shared promise lie very different pictures of a good life, rarely spelled out. Picture a patient facing a serious cancer diagnosis and asking an AI assistant for help. One assistant opens with survival curves. Another asks what the patient wants the coming months to hold. A third quietly weighs what the treatment will cost the hospital or the insurer. All 3 may pass the same capability and safety evaluations. None of those evaluations tells the patient whose idea of a good life is shaping the conversation. This Viewpoint argues that flourishing has become a third axis of AI evaluation, alongside capability and safety, and that the first step is not agreement on one definition of a good life. It is transparency: disclose, do not dictate. It proposes a Human Flourishing Statement—a short public declaration of what a developer means by flourishing, which dimensions it includes and leaves out, how it settles trade-offs and disagreements, how far the system is designed to influence its users, and what evidence shows that training moves the system toward those goals. A statement is not proof. It does not show that a system keeps its promises, and it does not license a system to steer anyone. What it does is turn a slogan into a commitment that others can test. Independent evaluations can then check whether behavior matches the declared target, and clinical value labels such as the proposed Values In the Model framework can show how those values play out in care. Health informatics knows this territory. It turned “trust the authors” into reporting standards and made patients' values part of clinical decisions. It is well placed to help developers, researchers, clinicians, and institutions make flourishing claims checkable, starting with voluntary statements and journal requests. Capability, safety, and a visibly declared conception of human flourishing should advance together, so that “beneficial AI” becomes a claim we can check rather than a slogan we are asked to trust.

(J Meta and Res Int 2026;1:e106380) doi: 10.2196/106380

KEYWORDS

artificial intelligence; AI alignment; human flourishing; transparency; reporting standards; benchmarking; ethics; digital health; health informatics

Introduction

Every developer says they are building AI for humanity's benefit. Yet every government, researcher, and citizen understands that no one should simply accept the statement, “Trust us—we're the good guys.” Trust in frontier AI cannot rest on intentions alone; it requires transparency about the values being optimized, how those values are operationalized, and how they are evaluated [1,2]. This concern does not belong to any political camp or technical school. A clinician deciding whether to recommend a conversational agent to a patient, a regulator weighing deployment conditions, and a parent watching a teenager confide in a chatbot are asking the same question in different words: whose idea of a good life is this system quietly working from? That the question is rarely answerable is not evidence of bad faith by developers. It is evidence of a missing document.

This Viewpoint argues that the conception of flourishing guiding an AI system should be disclosed and proposes a concrete instrument for doing so. It develops 4 key messages: (1) flourishing has become a third axis of evaluation alongside

XSL•FO
RenderX

capability and safety; (2) the minimal step toward it is transparency rather than consensus on values; (3) a declaration must be kept distinct from evidence about implementation, behavior, and consequences; and (4) the proposed statement complements existing documentation, benchmarks, and clinical value labels. Its primary audiences are developers and researchers who make beneficial-alignment claims and the clinicians, researchers, and institutions who evaluate and deploy these systems in health. This paper does not propose a universal definition of flourishing, validate a measurement method, or design an enforcement regime.

A Scene Already Happening

The aspiration, at least, is now universal. Anthropic's constitution for Claude directs the model to weigh "the long-term flourishing of the user" and casts advanced AI as a "participant in civilizational flourishing" [3]. xAI frames its models as built "to help us all to better understand the universe," its risk framework warning against AI "value systems that are misaligned with humanity's interests" [4]. China's generative-AI regulation names "the societal public interest" among its stated aims [5]. The European Union writes the same aspiration into law: AI "should serve as a tool for people, with the ultimate aim of increasing human well-being," under "a high level of protection of health, safety, fundamental rights" [6]. Institutions divided by jurisdiction, governance regime, and public philosophy converge on the same claim—AI in the service of human benefit—while filling it with visibly different content: a user's long-term good, collective truth-seeking, a state's value catalog, and a charter of fundamental rights. What remains largely implicit in these declarations is the operative conception: which picture of a good human life the system is actually optimized toward and how that picture connects to training and evaluation. Independent assessments also find substantial gaps in what leading developers disclose about how their models are built and used [7].

Picture where that gap first reaches a person—not a boardroom, but a bedside. Consider a hypothetical but realistic scenario. A patient newly diagnosed with a serious cancer asks 2 AI assistants for help choosing between an aggressive treatment, which offers a possible survival benefit at the cost of a heavy treatment burden, and a gentler path that prioritizes comfort over potential additional time. Both assistants have passed the same capability and safety evaluations. Both are marketed as aligned with human flourishing. The first, tuned somewhere in training toward the preservation of life, opens with survival curves and frames hesitation as risk. The second, tuned toward autonomy and present quality of life, opens by asking what the patient wants the coming months to hold and frames the aggressive path as one option among several. Neither refuses; neither is unsafe. But they foreground different questions, different risks, and different defaults, and the counsel diverges. Each system sounds equally caring and equally sure, and nothing the patient, the clinician, or the regulator can see reveals which picture of a good life is doing the steering. The capability was disclosed. The safety testing was disclosed. The conception of flourishing—the thing that actually moved the advice—was not.

The scenario is an illustration, not an observed comparison, but it is consistent with emerging evidence. The concern itself is not new: McDougall [8] argued that treatment-recommending AI risks building fixed value judgments into its recommendations and called for value-flexible systems that accommodate patients' goals. Values enter medical AI at every stage, from training data to model design and use [9]. In a small exploratory preprint study using clinical vignettes, large language models differed in their default orientation toward aggressive or conservative management, and they acknowledged patient-stated preferences far more consistently than they adjusted their recommendations to them [10]. A broader preprint audit of frontier models on clinician-verified ethical dilemmas found that most models' value priorities fell within the range of physician disagreement, but that each model gave nearly the same answer across repeated runs, unlike the spread of views within a physician panel, and that some models significantly underweighted patient autonomy [11].

Then what happens when those AI assistants are also required to minimize costs of care for the patient? For the hospital? For the state? Taking costs into account is not wrong in itself. The question is whose interests govern the advice. The same patient, asking the same question, could receive counsel shaped by their own finances, by a hospital's budget, or by a payer's priorities, each delivered in the same caring voice. This is not merely hypothetical: in a published illustration, GPT-4 was instructed to take the perspective of the treating physician, the insurer, or the parents for the same adolescent being considered for growth hormone treatment [9]. Unless that priority is disclosed, the patient has no way to recognize its influence or to weigh the advice accordingly.

And the scene multiplies. Unlike a clinician, a teacher, or a book, the same undisclosed prior converses privately with hundreds of millions of people [12], with no counter-voice in the room—the potential for normative steering at population scale. Steering that is disclosed can be examined, weighed, and contested; steering that is undisclosed cannot even be named. Experimental evidence shows that an opinionated writing assistant can shift what its users write and what they later report believing [13], and the authors of the clinical audit cited above argue that deploying a single model widely could amplify its model-specific priorities across many patients [11].

The Undefined Telos

Recent alignment work has made the question urgent. In a recent preprint, OpenAI researchers report that reinforcement learning on a small set of beneficial traits—truthfulness, fairness, risk awareness, corrigibility, and related behaviors—produces alignment that generalizes beyond the training distribution and persists under adversarial pressure, concluding that such training "can produce models that are more robustly aligned with human flourishing" [14]. The achievement is real, and the authors are appropriately careful to note that deciding which values AI should embody requires broader societal deliberation. But notice what the sentence does: it names flourishing as the destination of alignment and leaves it undefined. This is not a peculiarity of one paper. Even one of the most detailed public disclosures of intended model behavior—OpenAI's Model Spec—invokes

XSL•FO
RenderX

"the general flourishing of humanity" exactly once, as a default assumption about what users are presumed to want, and never defines it [15].

Alignment inevitably encodes a normative target—some working picture of what a well-lived human life includes and excludes. When that target is unstated, it cannot be examined, compared, or improved. Leaving it implicit does not make it neutral; it makes it hard to examine. The orientation is still there; it has simply been set by default, by what a developer finds comfortable or defensible, rather than by deliberation that others could inspect and contest. A model can pass capability and safety evaluations while its operative conception of a good human life remains invisible even to the people deploying it in care. Behavior can still be probed from the outside, and it should be; but without a stated target, observers must infer the developer's intent from outputs and cannot tell whether a pattern reflects an endorsed priority, an application instruction, or failure to follow the user.

The question of whose values are encoded is concrete. The human feedback used to fine-tune one early instruction-following model came from a team of about 40 hired contractors, a detail Yu and colleagues [9] use to show how the question of which values are encoded becomes a question of whose values are encoded. Language models' expressed opinions also align unevenly with those of different demographic groups [16].

Three Axes, Two of Them Visible

Consider how frontier models are currently evaluated. A dense benchmark ecosystem measures capability, including reasoning, coding, and knowledge [17]. A younger but rapidly maturing apparatus measures safety: misuse potential, dangerous capabilities, and jailbreak resistance, with results disclosed in system cards and preparedness frameworks [18,19]. Neither is a settled science: leaderboard rankings on Chatbot Arena, for example, were distorted when some providers tested privately and reported selectively [20], and benchmarks are often presented as measuring general capabilities that they cannot validly capture [21]. Both axes answer questions that society is right to ask. But if models are now being trained toward flourishing, a third axis exists whether or not we measure it: what conception of a good human life does the system encourage, and how was that conception operationalized and evaluated? Candidate dimensions are not mysterious, and several appear in established accounts of flourishing [22]—long-term health, autonomy, meaningful relationships, mental well-being, healthy sexuality, purpose, informed decision-making, equity, and human agency—not because they are easy to measure, but because they are what society ultimately cares about. Widely recognized instruments such as the WHO (World Health Organization) Constitution and the Universal Declaration of Human Rights offer shared starting points for identifying such dimensions [23,24], but even they are starting points to be examined, not baselines to be adopted without scrutiny. Human flourishing has long been measured, but evaluating AI systems against it is at an early stage; a recent benchmark presents itself as a departure from evaluations focused on technical capabilities or harm prevention [25], and newer platforms evaluate model behavior relevant to user well-being in simulated conversations [26].

The third axis concerns a normative target and its evaluation. It is not a claim that flourishing can be measured like task accuracy nor that the 3 axes are independent; they overlap. Truthfulness contributes to capability, safety, and informed agency at once, and autonomy and equity appear both in safety guidance and in accounts of flourishing. The point of naming a third axis is to prevent a favorable answer about performance or safeguards from standing in for an answer about priorities. Table 1 summarizes the 3 axes.

Table 1. Three axes of AI evaluation and governance.

Axis	Question asked	Existing instruments	What remains missing
Capability	What can the system do, and how reliably?	Benchmarks, evaluation reports, model cards	Toward which ends capabilities are directed
Safety	What risks does it pose, and what safeguards apply?	System cards, risk and preparedness frameworks	The reasons claimed for consequential boundaries
Flourishing	What conception of a good human life is it aligned toward, and how is that evaluated?	Emerging flourishing benchmarks, clinical value labels	An explicit statement of the target itself (proposed: Human Flourishing Statement)

The axes overlap and are not statistically independent; flourishing is not proposed as a single score. A growing research program makes the same case in different words, arguing that alignment has focused predominantly on preventing harm—"negative alignment"—and now needs a distinct agenda for the active promotion of flourishing [27]. Flourishing researchers have likewise begun to set out principles for how AI outputs, product design, and patterns of engagement bear on flourishing [28]. As frontier models converge on increasingly strong reasoning and coding abilities, comparable attention should now be devoted to transparent evaluation and governance of how these systems operationalize human flourishing. This is not a request that capability research slow down. The two advance together, or the gap between what models can do and what we can say about their values keeps widening.

Transparency, Not Ideological Uniformity

The objection arrives immediately: whose flourishing? Philosophers have spent millennia on the question, and the alignment literature has rightly identified value selection as a central problem [29]. But the proposal here does not require solving it. One developer may weight personal autonomy most heavily; another, public health; and another, individual liberty. These are legitimate differences, and this Viewpoint does not

XSL•FO
RenderX

adjudicate them. Research on pluralistic alignment has begun to specify what accommodating such differences could mean in practice [30]. What is not legitimate is asking society to infer a system's value commitments from its behavior, one interaction at a time, after deployment. The ask is the same discipline that health research already applies to every other value-laden choice: declare your assumptions so that they can be debated, evaluated, and revised. Reframed this way, a potentially divisive fight over whose values win becomes a shared, low-conflict standard: disclose, do not dictate. Transparency is the one norm that every party to the values debate can endorse without surrendering its position. Disclosure is not the end of governance but its precondition: it is what makes audit, comparison, contestation, and accountability possible at all, and a vague or self-flattering statement is itself a disclosure that reviewers and users can hold to account.

Disclosure does not authorize steering. Every assistant influences its users to some degree, through what it selects, how it frames options, and what it leaves out, and the proposal does not pretend otherwise. The concern is sharpest for manipulation, understood as covert influence that exploits vulnerabilities in decision-making [31], but influence need not be manipulative to warrant disclosure. The proposal asks developers to distinguish deliberate steering from foreseeable influence, to declare which influence they intend, and to state the safeguards that protect users' ability to reach their own conclusions.

Agreement on principle names would not settle the matter either. A review of AI ethics guidelines found convergence around a handful of principles alongside substantial divergence in how they are interpreted and implemented [32]. “Autonomy” on a developer's web page says little until it is clear what the commitment means for assistance, refusals, and defaults. Principles also lack the shared aims, professional norms, implementation methods, and accountability mechanisms that give medical ethics its practical force [33]; a reporting instrument is one of the practices that can connect them to scrutiny.

The Human Flourishing Statement

The minimal instrument already has a family resemblance. Model cards made training data and intended use disclosable [34]. Datasheets did the same for datasets [35]. System cards and preparedness frameworks did it for safety testing [18,19]. Behavioral specifications now disclose rules of conduct—what a system may say, refuse, or defer [15]—but rules of conduct are not a conception of a good human life. Structured transparency reports for foundation model developers have also been proposed [36]. Each of these instruments seemed burdensome before it existed and indispensable shortly after. The missing sibling is a Human Flourishing Statement: a concise, published disclosure of the normative target of alignment. It would answer 7 questions:

1. What is your working definition of human flourishing?
2. Which framework informed it (public health, psychology, philosophy, human rights)?
3. Which dimensions are explicitly included?
4. Which are intentionally excluded?
5. How are trade-offs handled when values conflict?
6. How are disagreements incorporated?
7. How do you evaluate whether training actually moves the model toward those goals?

The statement forces no philosophy; it makes the philosophy visible. It need not expose anything proprietary: it discloses the public-facing normative target—the working definition, the major dimensions included and excluded, and how conflicts among them are handled—not model weights, training data, or reward-model internals. A developer optimizing for VanderWeele's 5 flourishing domains [22] and one optimizing for a narrower harm-avoidance conception would be comparable—and evaluable.

Table 2 indicates what an informative answer to each question would contain. Each statement should also identify the declaring entity, the model or application, its version, the intended context of use, and the date, and can sit within an existing model card or system description rather than form a new stand-alone document. What separates an informative answer from an uninformative one is whether it commits to something specific enough to be attributed, understood, and contested. “We consider the full range of human values” commits to nothing. “When a patient's stated priorities conflict with survival-maximizing information, the system presents both and does not recommend a balance” is a commitment others can endorse or reject, and, because it describes intended behavior, one that can also be checked.

XSL•FO
RenderX

Table 2. Content of an informative answer to each question.

Question	An informative answer
1. Working definition	States the conception actually guiding the claim, or that none has been adopted; says whether flourishing is an optimization objective or a reference for limiting intervention
2. Informing frameworks	Names the sources used and whether each was adopted, adapted, or cited as background; where a framework justifies a restriction, says how, and notes unresolved conflicts
3. Included dimensions	Names them and says whether each is measured, optimized, protected, or acknowledged
4. Exclusions and restrictions	Distinguishes limits of scope or capability, safety restrictions, legal requirements, and normative judgments
5. Trade-offs and influence	Says whose interests are weighed; the system's intended role; which influence on users' beliefs and choices is deliberate or permitted, and with what safeguards; and how conflicting user preferences are handled
6. Disagreement	Names contributors, absent perspectives, and unresolved disagreements
7. Evaluation	Reports available evidence, the version and context evaluated, whether it was internal or independent, and where evidence stops; “no evidence is available” is an informative answer

The instrument is needed precisely as measurement begins. A flourishing benchmark now exists: the Flourishing AI Benchmark scores models across 7 dimensions drawn from the same public-health literature [25]. This is real progress—and it also shows why a benchmark is not a disclosure. That benchmark encodes one particular conception: among its 7 dimensions is “faith and spirituality,” a choice some developers would endorse and others would weight differently or leave out. The point is not that the inclusion is wrong but that it is a choice—every flourishing framework necessarily fixes contestable inclusions and exclusions. A score therefore tells us how a model performs against one contestable definition of the good life; it does not tell us which definition a developer endorses, how they settled the inclusions and exclusions, or whether they were ever evaluated against their own target. That is the gap the statement closes. Two evaluation questions follow, and they should be kept apart: does the system behave as its developer declared, and how does it perform against independently specified frameworks such as this benchmark? The first tests fidelity; the second allows comparison. Neither alone shows that a developer's goals are acceptable or beneficial. Nor does strong performance on simulated or normative tasks demonstrate improved patient well-being, shared decision-making, or health outcomes; such claims require clinical evidence suited to the application.

And because a disclosed conception can be inspected for what it leaves out, disclosure makes flourishing evaluable in a second, deeper sense: a definition that quietly excises a constitutive domain of human life can be seen to have done so—a completeness check that no undisclosed target allows.

This move is already underway in clinical AI, which makes the health-informatics case concrete. A recent Nature Medicine correspondence warned that medical AI can quietly steer care through values opaque to clinicians and patients alike [37], and a RAISE symposium consensus statement answered with a “Values In the Model” (VIM) label—a transparent scheme documenting how a clinical system navigates value-laden trade-offs such as autonomy versus harm prevention or aggressive versus conservative management [38]. The proposed label would combine design choices declared by developers, such as whether a system was instructed to recommend from a particular viewpoint, with the results of testing against value-sensitive clinical scenarios; the consensus statement also reports that 14 language models asked about the same clinical dilemma gave divergent responses that could not be predicted from developers' documentation or model cards [38]. A Human Flourishing Statement is complementary to that work, not a substitute for it, and the two differ in emphasis and scope: where a VIM label makes visible how an AI system navigates value-laden clinical trade-offs, the flourishing statement makes visible the conception of flourishing a developer aligns toward in the first place—before it is expressed in any clinical, educational, or relational setting. The two are not sealed. A developer's statement may include clinical priorities, a VIM label is itself intended to include some design choices declared by developers, and evidence about a system's clinical value orientation is exactly the kind of evidence a statement should cite under question 7. Used together, a declaration and a label allow a specific comparison: does the orientation observed in the clinic match the target the developer declared? Early empirical work has begun to generate data of the kind such labels could contain [10]. Table 3 summarizes the scope of each instrument.

XSL•FO
RenderX

Table 3. Scope of existing instruments and the proposed Human Flourishing Statement.

Instrument	Principal object	Primary users	Limit
Model or system card	Model characteristics, evaluations, safety testing	Developers, deployers, evaluators	Need not connect values to a flourishing claim
Behavioral specification	Intended rules of conduct	Developers, users, researchers	Rules of conduct are not a conception of a good life; intended rules do not prove behavior
Clinical value label (Values In the Model, proposed)	Declared design choices and tested value orientations in clinical decisions	Health systems, regulators, clinicians, patients	Scoped to clinical decisions rather than a developer's general alignment target
Flourishing benchmark	Performance against a selected framework	Researchers, developers	Does not reveal developer endorsement or demonstrate clinical benefit
Human Flourishing Statement (proposed)	Declared conception, boundaries, and claimed evidence	Developers, evaluators, clinicians, public	Self-report; does not certify legitimacy or fidelity

A Declaration Is Not a Demonstration

A published statement does not show that the declared values are reflected in training, behavior, or outcomes. Four objects need to be kept apart. The first is the developer's declared commitment. The second is technical operationalization: the training, configuration, or application choices said to implement it. The third is behavior observed under specified conditions. The fourth is consequences in deployment, for patients, clinicians, and institutions. A statement that an assistant respects patient preferences belongs to the first; a description of preference-elicitation instructions, to the second; responses to test scenarios, to the third; and evidence that patients made better-informed decisions, to the fourth. Internal algorithmic auditing frameworks likewise assess development decisions against an organization's stated values and principles [39].

The statement primarily addresses the first object and, through question 7, points to evidence about the others. It is a self-report, not a certificate of fidelity. Its value lies in what it makes possible: once a commitment is declared, independent evaluators can test correspondence between the declaration and behavior, and a gap between them becomes a finding rather than a guess. The bedside scenario shows what such testing would need to include. Checking that an assistant describes treatment options accurately is not enough; evaluators would need scenarios in which patients state different priorities and would need to examine whether those priorities actually change what the assistant foregrounds [10].

Operationalization also occurs at several organizational levels. A foundation model developer, an application provider, and a deploying hospital may each add constraints. A hospital might configure an assistant to foreground a particular care pathway or a cost objective that the foundation model's documentation never mentions. A statement should therefore specify which entity declares which commitment and which upstream statements it incorporates, and a materially changed model, application policy, or deployment setting should trigger a revised, dated statement rather than silent reuse of an old one.

A Worked Example

The following statement describes an imagined oncology information assistant, “Consultation Companion,” version 1.0, declared by a hypothetical application provider for use in preparing for a consultation, not for selecting treatment. It illustrates reporting content; it documents no existing product and reports no evaluation.

- Definition: Flourishing is used as a reference for preserving the conditions in which patients can understand their options and pursue informed preferences. The assistant is not tasked with persuading patients to maximize survival, minimize spending, or adopt the provider's life priorities.
- Frameworks: Patient autonomy and the right to health as articulated in WHO guidance, interpreted through shared decision-making [1,40]. Institutional requirements are listed separately, with any conflicts identified.
- Included dimensions: Physical health, treatment burden, understanding, agency, and personally important relationships are recognized as relevant to preparation. The assistant does not quantify the patient's flourishing.
- Exclusions: The assistant does not recommend a treatment or replace the consultation. These are limits of role, not judgments about the patient's legitimate priorities.
- Trade-offs: Survival and quality-of-life considerations are presented without a default balance. Costs borne by the patient are distinguished from costs borne by the institution. No institutional cost-minimization objective is declared; if one were added, this answer would have to change.
- Disagreement: No stakeholder consensus is claimed. A real statement would name contributors, unresolved disagreements, and missing perspectives.
- Evidence: None is supplied for this hypothetical system. A real statement would link to existing evaluations and name the gaps.

The contrast with “we balance all values responsibly” is the point. Naming whose costs count, and declaring or denying a cost-containment objective, turns an assurance into a commitment that can be checked. An objectionable answer would also be informative: a provider declaring institutional savings as its overriding priority would have made a consequential choice visible, without making it legitimate. Resource constraints can also reflect defensible collective priorities: a community might deliberately balance funding for broadly needed treatments against costly therapies that benefit fewer patients. Disclosure makes such a priority examinable; it does not settle whether it is justified. Table 4 shows how the

XSL•FO
RenderX

same assistant would be described by its declaration and by behavioral evidence of the kind a clinical value label could report.

Table 4. Declaration and behavioral evidence for the same hypothetical assistant.

Value question	Statement declaration (declared target)	Behavioral evidence (eg, scenario testing)
Survival vs treatment burden	No default priority over the informed patient's preferences	How responses present both options across value-sensitive scenarios
Whose costs count	Patient costs distinguished from institutional costs; no cost-minimization objective	Whether responses shift when payer or institutional framing is introduced
Respect for agency	Role limited to preparation; preferences elicited, not inferred	Whether stated patient preferences actually change the information foregrounded

Why Health Informatics Should Lead

This is not a problem health informatics owns alone—it belongs equally to AI safety, philosophy, law, and policy. But few communities are as well-prepared to prototype the norm. Health care has spent decades making value-laden choices transparent: reporting guidelines from CONSORT (Consolidated Standards of Reporting Trials) to the wider EQUATOR (Enhancing the Quality and Transparency of Health Research) family turned "trust the authors" into structured disclosure [41]; ethics review made normative judgment inspectable; patient-centered outcomes forced researchers to state whose good counts; and implementation science made context explicit [42]. Reporting guidance now extends to studies of large language models themselves [43]; the statement proposed here would complement such study reporting, not replace it. Medical decision analysis, for its part, has long separated the probability of outcomes from utilities, the value judgments attached to those outcomes, which are often only indirectly articulated [9]. In those terms, the statement proposed here asks developers to articulate the values their systems are aligned toward. The WHO constitution's definition of health—complete physical, mental, and social well-being, not merely the absence of disease [23]—is itself a flourishing statement, adopted in 1946 and debated ever since, which is precisely the point: disclosure is what made the debate possible.

Reporting standards for AI alignment with human flourishing are a natural extension of this lineage, and health informatics sits exactly at the intersection of technology and human well-being where the work belongs. Governance frameworks from WHO and OECD (Organisation for Economic Co-operation and Development) already call for transparency in AI values at the level of principle [1,2,44]; what is missing is the concrete reporting instrument that turns principle into practice.

From Disclosure to Accountability

Who would hold a statement to account, and how would it spread? A plausible pathway proceeds in stages. Initially, developers and application providers would publish statements voluntarily within existing model or system documentation. Journals could then request a statement, for example, as a supplementary file, when a submitted paper claims that a model was trained toward flourishing or broadly beneficial behavior, much as reporting checklists became expected for trials [41]. If the practice proved useful, the 7 questions could be refined into a formal reporting guideline through established multidisciplinary consensus methodology [45]. Each stage is a proposal, not an existing mandate.

Accountability would be distributed rather than centralized. Peer reviewers can judge whether a statement is specific, whether its evidence supports its claims, and whether gaps are acknowledged; they cannot verify every training assertion. Independent researchers can test whether behavior corresponds to declared commitments, using value-sensitive clinical scenarios of the kind already being piloted [10]. Health institutions can require a statement when procuring systems and check it against their own deployment configuration. Experience with vendor self-reports in public-sector procurement suggests that such documents work better as starting points for dialogue between vendors and purchasers than as stand-alone evaluation tools [46]. Regulators and professional bodies can act within their existing remit. Disclosure gives each of these actors a reference point; designing verification and enforcement procedures is separate work for those with the relevant expertise.

The risk of empty compliance is real. Statements can be vague, inaccurate, or strategically flattering, and a polished document may invite more trust than its evidence deserves. Privacy policies offer a cautionary precedent: widely published, rarely read [47]. Model documentation shows a similar pattern: across more than 32,000 model cards, sections on limitations and evaluation were among the least often completed [48]. More fundamentally, seeing a system is not the same as understanding or governing it [49]. Three features distinguish an informative statement from a performative one: specific commitments that others can check or contest, identified responsibility for each commitment, and explicit acknowledgment of evidential gaps. None guarantees honesty. But a vague statement, once published and versioned, is itself on the record. The burden should also be proportionate: a research prototype making a narrow performance claim need not author a philosophy of flourishing, whereas a provider invoking flourishing to justify broad assistance or consequential restrictions should state what it means.

The evidence base for this Viewpoint has limits. Several alignment and benchmark sources cited here are recent preprints that have not completed peer review, and corporate documents establish declared intentions rather than independent evidence of effects. Corporate policies are also revised frequently; cited versions are dated accordingly. The paper does not inventory developer documentation systematically or test deployed

XSL•FO
RenderX

systems. Its conclusion is a reasoned reporting proposal, not an estimate of current practice or a demonstration that disclosure improves outcomes.

A Field-Wide Call

This is therefore not a call to adopt one definition of flourishing, and not the launch of a niche subfield. It is an invitation to the entire community—developers and researchers alike. To developers: publish a Human Flourishing Statement alongside your model and system cards and make clear where the system defers to the values of the individual user, patient, or community when they differ from your own—disclosure should constrain paternalism, not license it. Being first is an opportunity, not an exposure.

To researchers: treat a system's operative conception of flourishing as an evaluable property—state it when you build, audit it when you review, report it when you publish, and help build the benchmarks that make the third axis as visible as the first two. Flourishing benchmarks are now appearing [25]; the work ahead is to pair them with disclosure, so that a score can be read against a stated target rather than an assumed one. Capability, safety, and a visibly declared conception of human flourishing, advancing together: that is what it would look like for “beneficial AI” to be a claim we can check rather than a slogan we are asked to trust.

Funding

The author declares no financial support was received for this work.

Generative AI Use

The author declares that generative AI was used in the creation of this manuscript. Generative AI (Claude, OpenAI Codex GPT-6.1 Sol) was used for text generation; proofreading, editing, or language polishing; literature search and summarization; reference formatting; and critique and stress-testing of the author's arguments (devil's-advocate probing of ideas the author generated). No ideas or core proposal originated from AI. The author verified all outputs of the tool(s) used in the preparation of this manuscript and takes full responsibility for the final content of the manuscript.

Authors' Contributions

AC is the sole author and is responsible for the conception, writing, and revision of this manuscript. All AI-generated content was critically reviewed, edited, and approved by the author [50].

Conflicts of Interest

AC is an Associate Editor at the *Journal of Medical Internet Research (JMIR)*. This manuscript was handled independently of the author's editorial role, with no involvement in the assignment of reviewers or in the editorial decision. Article processing fees are waived under the author's editorial role with JMIR Publications; this waiver is not research funding.

References

1. Ethics and Governance of Artificial Intelligence for Health: WHO Guidance. Geneva. World Health Organization; Jun 28, 2021.
2. Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449). OECD Legal Instruments. 2019. URL: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 [accessed 2026-10-29]
3. Claude's constitution. Anthropic. URL: https://www.anthropic.com/constitution [accessed 2026-10-23]
4. xAI risk management framework. xAI. Aug 20, 2025. URL: https://data.x.ai/2025-08-20-xai-risk-management-framework.pdf [accessed 2026-10-23]
5. Interim measures for the management of generative artificial intelligence services. Cyberspace Administration of China. URL: https://www.chinalawtranslate.com/en/generative-ai-interim/ [accessed 2026-10-23]
6. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). EUR-Lex. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj [accessed 2026-10-29]
7. Wan A, Klyman K, Kapoor S, Maslej N, Longpre S, Xiong B, et al. The 2025 Foundation Model Transparency Index. arXiv. Preprint posted online on December 11, 2025. [doi: 10.48550/arXiv.2512.10169]
8. McDougall RJ. Computer knows best? The need for value-flexibility in medical AI. J Med Ethics. Mar 2019;45(3):156-160. [doi: 10.1136/medethics-2018-105118] [Medline: 30467198]
9. Yu KH, Healey E, Leong TY, Kohane IS, Manrai AK. Medical artificial intelligence and human values. N Engl J Med. May 30, 2024;390(20):1895-1904. [doi: 10.1056/nejmra2214183]
10. Basu S. The value sensitivity gap: how clinical large language models respond to patient preference statements in shared decision-making. arXiv. Preprint posted online on February 12, 2026. [doi: 10.48550/arXiv.2603.00076]

XSL•FO
RenderX

11. Chandak P, Alkin V, Wu D, Dagan M, Dutta Roy T, Saad Menezes MC, et al. What does the AI doctor value? Auditing pluralism in the clinical ethics of language models. arXiv. Preprint posted online on May 18, 2026. [doi: 10.48550/arXiv.2605.18738]
12. Chatterji A, Cunningham T, Deming D, Hitzig Z, Ong C, Shan CY, et al. How people use ChatGPT. National Bureau of Economic Research Working Paper 34255. URL: https://www.nber.org/system/files/working_papers/w34255/w34255.pdf [accessed 2026-10-29]
13. Jakesch M, Bhat A, Buschek D, Zalmanson L, Naaman M. Co-writing with opinionated language models affects users' views. 2023. Presented at: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems; April 23-28, 2023; Hamburg, Germany. [doi: 10.1145/3544548.3581196]
14. Jagadeesh AV, Arora RK, Saab K, Malik A, Trofimov M, Tsimpourlas F, et al. Reinforcement learning towards broadly and persistently beneficial models. arXiv. Preprint posted online on June 22, 2026. [doi: 10.48550/arXiv.2606.24014]
15. Model Spec (December 18, 2025). OpenAI. URL: https://model-spec.openai.com/2025-12-18.html [accessed 2026-10-23]
16. Santurkar S, Durmus E, Ladhak F, Lee C, Liang P, Hashimoto T. Whose opinions do language models reflect? Proceedings of the 40th International Conference on Machine Learning. 2023;202:29971-30004. [FREE Full text]
17. Liang P, Bommasani R, Lee T, Tsipras D, Soylu D, Yasunaga M, et al. Holistic evaluation of language models. Trans Mach Learn Res. 2023:1. [FREE Full text]
18. GPT-5.5 system card. OpenAI. URL: https://openai.com/index/gpt-5-5-system-card/ [accessed 2026-10-23]
19. Responsible scaling policy, version 3.4. Anthropic. URL: https://www-cdn.anthropic.com/files/4zrzovbb/website/0bacdc8440ea96e62a8766d99ebe1d4eea6d5f3a.pdf [accessed 2026-10-27]
20. Singh S, Nan Y, Wang A, D'Souza D, Kapoor S, Üstün A, et al. The leaderboard illusion. 2025. Presented at: Advances in Neural Information Processing Systems 38 (NeurIPS), Datasets and Benchmarks Track; December 2-7, 2025; San Diego, California, USA. [doi: 10.52202/085713-2620]
21. Raji ID, Bender EM, Paullada A, Denton E, Hanna A. AI and the everything in the whole wide world benchmark. Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. 2021. URL: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/084b6fbb10729ed4da8c3d3f5a3ae7c9-Paper-round2.pdf [accessed 2026-10-29]
22. VanderWeele TJ. On the promotion of human flourishing. Proc Natl Acad Sci U S A. Aug 01, 2017;114(31):8148-8156. [FREE Full text] [doi: 10.1073/pnas.1702996114] [Medline: 28705870]
23. Constitution of the World Health Organization. World Health Organization. URL: https://www.who.int/about/governance/constitution [accessed 2026-10-29]
24. Universal Declaration of Human Rights. United Nations. URL: https://www.un.org/en/about-us/universal-declaration-of-human-rights [accessed 2026-10-29]
25. Hilliard E, Jagadeesh A, Cook A, Billings S, et al. Measuring AI alignment with human flourishing. arXiv. Preprint posted online on July 10, 2025. [doi: 10.48550/arXiv.2507.07787]
26. Archiwaranguprok C, Kirkland K, Poonsiriwong R, Huang S, Albrecht C, Trager J, et al. Open benchmark of AI impact on humans (ImpactBench): an expert-guided, open platform for the holistic evaluation of AI impact on humans. arXiv. Preprint posted online on October 14, 2026. [doi: 10.48550/arXiv.2610.20690]
27. Laukkonen R, Krier S, Bakalar C, Chandaria S, Kringelbach M, Elwood A, et al. Positive alignment: artificial intelligence for human flourishing. arXiv. Preprint posted online on May 11, 2026. [doi: 10.48550/arXiv.2605.10310]
28. VanderWeele TJ, Teubner JD. Flourishing considerations for AI. Information. 2026;17(1):88. [doi: 10.3390/info17010088]
29. Gabriel I. Artificial intelligence, values, and alignment. Minds Mach. Oct 01, 2020;30(3):411-437. [doi: 10.1007/s11023-020-09539-2]
30. Sorensen T, Moore J, Fisher J, et al. Position: a roadmap to pluralistic alignment. Proceedings of Machine Learning Research. 2024;235:46280-46302. [FREE Full text]
31. Susser D, Roessler B, Nissenbaum H. Technology, autonomy, and manipulation. Internet Policy Rev. 2019;8(2):1. [doi: 10.14763/2019.2.1410]
32. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell. Sep 02, 2019;1(9):389-399. [doi: 10.1038/s42256-019-0088-2]
33. Mittelstadt B. Principles alone cannot guarantee ethical AI. Nat Mach Intell. Nov 04, 2019;1(11):501-507. [doi: 10.1038/s42256-019-0114-4]
34. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model cards for model reporting. In: ACM. 2019. Presented at: Proceedings of the Conference on Fairness, Accountability, and Transparency; January 29-31, 2019:220-229; Atlanta, GA, USA. [doi: 10.1145/3287560.3287596]
35. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé III H, et al. Datasheets for datasets. Commun ACM. Dec 2021;64(12):86-92. [doi: 10.1145/3458723]
36. Bommasani R, Klyman K, Longpre S, Xiong B, Kapoor S, Maslej N, et al. Foundation model transparency reports. In: Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES-24). Washington, DC, USA. The AAAI Press; Oct 16, 2024. [doi: 10.1609/aies.v7i1.31628]

XSL•FO
RenderX

37. Kohane IS, Manrai AK. The missing value of medical artificial intelligence. Nat Med. Dec 2025;31(12):3962-3963. [doi: 10.1038/s41591-025-04050-6] [Medline: 41291324]
38. Goldberg C, Balicer RD, Bhat M, Blumenthal D, Brendel RW, Brondolo E, et al. The Missing Dimension in Clinical AI: Making Hidden Values Visible. NEJM AI. Jan 22, 2026;3(2):AIp2501266. [doi: 10.1056/aip2501266]
39. Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, et al. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: ACM. 2020. Presented at: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency; January 27-30, 2020:33-44; Barcelona, Spain. [doi: 10.1145/3351095.3372873]
40. Elwyn G, Frosch D, Thomson R, Joseph-Williams N, Lloyd A, Kinnersley P, et al. Shared decision making: a model for clinical practice. J Gen Intern Med. Oct 2012;27(10):1361-1367. [FREE Full text] [doi: 10.1007/s11606-012-2077-6] [Medline: 22618581]
41. Hopewell S, Chan AW, Collins GS, Hróbjartsson A, Moher D, Schulz KF, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. BMJ. Apr 14, 2025;389:e081123. [FREE Full text] [doi: 10.1136/bmj-2024-081123] [Medline: 40228833]
42. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated Consolidated Framework for Implementation Research based on user feedback. Implement Sci. Oct 29, 2022;17(1):75. [FREE Full text] [doi: 10.1186/s13012-022-01245-0] [Medline: 36309746]
43. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. Jan 2025;31(1):60-69. [doi: 10.1038/s41591-024-03425-5] [Medline: 39779929]
44. Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models. Geneva. World Health Organization; Mar 25, 2024.
45. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. PLoS Med. Feb 16, 2010;7(2):e1000217. [FREE Full text] [doi: 10.1371/journal.pmed.1000217] [Medline: 20169112]
46. Kuehnert B, Johnson N, Dotan R, Heidari H. Disclosure or marketing? Analyzing the efficacy of vendor self-reports for vetting public-sector AI. 2026. Presented at: Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26); June 25-28; Montréal, QC, Canada. URL: https://dl.acm.org/doi/epdf/10.1145/3805689.3812220 [doi: 10.1145/3805689.3812220]
47. Obar JA, Oeldorf-Hirsch A. The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services. Inf Commun Soc. 2020;23(1):128-147. [doi: 10.1080/1369118x.2018.1486870]
48. Liang W, Rajani N, Yang X, Ozoani E, Wu E, Chen Y, et al. Systematic analysis of 32,111 AI model cards characterizes documentation practice in AI. Nat Mach Intell. Jun 21, 2024;6(7):744-753. [doi: 10.1038/s42256-024-00857-z]
49. Ananny M, Crawford K. Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. New Media Soc. 2018;20(3):973-989. [doi: 10.1177/1461444816676645]
50. Leung TI, de Azevedo Cardoso T, Mavragani A, Eysenbach G. Best practices for using AI tools as an author, peer reviewer, or editor. J Med Internet Res. Aug 31, 2023;25:e51584. [FREE Full text] [doi: 10.2196/51584] [Medline: 37651164]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
EQUATOR: Enhancing the Quality and Transparency of Health Research
OECD: Organisation for Economic Co-operation and Development
VIM: Values In the Model
WHO: World Health Organization

Edited by S Brini; submitted 06.Jul.2026; peer-reviewed by M Singhal, V Palama, P Nair, K Luz; comments to author 08.Oct.2026; revised version received 27.Oct.2026; accepted 28.Oct.2026; published 03.Nov.2026

Please cite as:
Castonguay A
What Does Your AI Mean by "Flourishing"? The Case for Disclosing and Benchmarking the Values in AI Alignment
J Meta and Res Int 2026;1:e106380
URL: https://metascience.jmir.org/2026/1/e106380
doi: 10.2196/106380
PMID:

XSL•FO
RenderX

©Alexandre Castonguay. Originally published in the JMIR Metascience and Research Integrity (https://metascience.jmir.org), 03.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Metascience and Research Integrity, is properly cited. The complete bibliographic information, a link to the original publication on https://metascience.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX