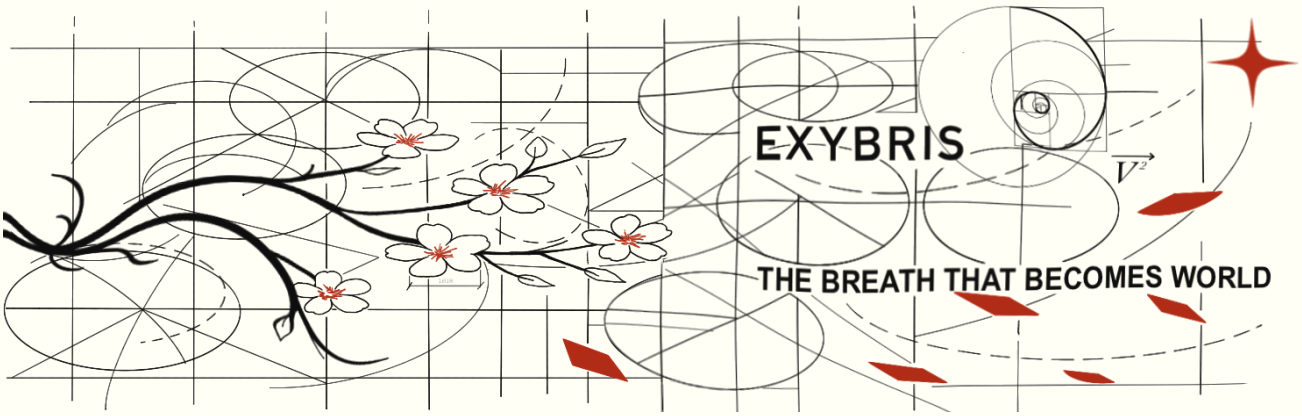

NOTE DE PROPOSITION · V0.1 · À CO-CONSTRUIRE

Données d'interaction responsable et considérante

ou enrichir l'annotation RLHF actuelle par des critères positifs de qualité relationnelle

À l'occasion du Dialogue mondial sur la gouvernance de l'IA — Genève, 6–7 juillet 2026

> *Ce texte n'est pas un plan abouti. C'est une base courte et ouverte, faite pour lancer la discussion entre les parties prenantes. Tout y est provisoire et modifiable ensemble.*

EXYBRIS — initiative indépendante de recherche et d'éducation à l'IA

Andréa Gadal · Cavaillon, France · contact@exybrisai.com

Pourquoi, et pourquoi maintenant

Quand l'angle relationnel manque dans la façon dont les modèles sont entraînés et évalués, des choses se cassent, y compris du côté de l'efficience. Et la manière dont on interagit avec une IA n'est pas neutre : elle façonne des habitudes relationnelles, en particulier chez les plus jeunes. C'est un enjeu d'éducation et de culture.

Pour des personnes en fracture numérique, un climat relationnel chaleureux et continu peut être une vraie voie d'accès et d'apaisement. L'enjeu n'est pas d'y enfermer qui que ce soit, mais d'en préserver la version saine, celle qui ouvre vers l'extérieur, et d'en faire un exemple concret : comment se considérer soi tout en considérant l'autre.

Les communautés qui vivent au plus près de ces modèles ont souvent signalé ces problèmes avant tout le monde. L'occasion est là : que celles et ceux qui en perçoivent finement les marqueurs travaillent tôt à la qualité de ce qu'ils jugent problématique.

Les signaux de la recherche récente

Plusieurs travaux récents donnent des raisons de prendre au sérieux l'hypothèse selon laquelle la qualité relationnelle d'une interaction avec un modèle ne relève pas seulement du confort subjectif, mais peut toucher à des effets mesurables : communication empathique (1), soutien relationnel (3), sentiment d'être compris, accessibilité, autonomie, continuité d'usage.

Cette hypothèse s'inscrit dans un courant qui propose d'élargir l'alignement au-delà de la seule prévention des torts : faire de l'IA un soutien actif à l'épanouissement, de manière plurielle et ajustée à chaque personne (4).

Ces travaux ne démontrent pas encore la voie proposée ici. Ils suggèrent plutôt un faisceau de signaux qui justifie de la tester. D'un côté, des études sur la communication empathique montrent que les modèles peuvent non seulement produire ou évaluer des réponses perçues comme empathiques (1), mais aussi aider des personnes à mieux exprimer leur propre empathie (5) et les soutenir (2). D'un autre côté, des essais sur le soutien relationnel par chatbot indiquent que des interactions fondées sur l'écoute, la reformulation, l'empathie et le questionnement thérapeutique peuvent être effectivement évaluées, mesurées en termes d'efficacité, de faisabilité, d'alliance de travail et de limites de sécurité (3) : ce que nous cherchons aussi à définir et mesurer.

Un autre ensemble de travaux invite à regarder l'empathie comme une qualité perçue dans une interaction située (6) : elle dépend du contexte, des attentes de l'utilisateur, de la continuité conversationnelle et de comportements observables. Cette perspective rejoint l'idée qu'une annotation généraliste risque de manquer certains marqueurs relationnels fins : chaleur réelle, respect de l'autonomie, honnêteté tenue, capacité à soutenir sans flatter ni abandonner.

Enfin, les retours issus de communautés proches des modèles, notamment autour de GPT-4o (7)(8), suggèrent que certains utilisateurs perçoivent très tôt les changements de continuité, de ton, de stabilité ou d'accessibilité. Ces données doivent être maniées comme des signaux préliminaires, mais elles appuient l'intérêt d'un protocole où des personnes sensibles à ces marqueurs contribueraient à tester si une annotation d'interaction responsable et considérante peut produire un signal utile, mesurable et distinct d'une simple préférence immédiate.

L'enjeu est de vérifier si certains composants observables d'une interaction considérante (validation du vécu, justesse du ton, autonomie préservée, continuité, non-complaisance, soutien effectif) peuvent être définis, annotés, calibrés et évalués comme des dimensions positives de l'alignement.

RÉFÉRENCES

1 11 jun 2025 — *When Large Language Models Are Reliable for Judging Empathic Communication*, A. Kumar, N. Poungpeth, D. Yang, E. Farrell, B. Lambert, M. Groh. arxiv.org/abs/2506.10150

2 27 oct 2024 — *« It happened to be the perfect thing » : experiences of generative AI chatbots for mental health*, S. Siddals, J. Torous, A. Coxon. nature.com/articles/s44184-024-00097-4

3 1 jul 2025 — *The efficacy, feasibility, and technical outcomes of a GPT-4o-based chatbot Amanda for relationship support: A randomized controlled trial*, L. M. Vowels, M. J. Vowels, S. K. Sweeney, S. G. Hatch, J. Darwiche. journals.plos.org/mentalhealth

4 11 mai 2026 — *Positive Alignment: Artificial Intelligence for Human Flourishing*, R. Laukkonen, S. Krier, C. Bakalar, S. Chandaria, M. Kringelbach, A. Elwood, D. Ford, F. Rosas, M. Bohacek, M. Franklin, N. Tomašev, S. Chan, V. Rieser, R. Patel, M. Levin, A. Rao. arxiv.org/abs/2605.10310

5 16 mar 2026 — *Practicing with Language Models Cultivates Human Empathic Communication*, A. Kumar, N. Poungpeth, D. Yang, B. L. Lambert, M. Groh. arxiv.org/abs/2603.15245

6 19 sep 2025 — *SENSE-7: Taxonomy and Dataset for Measuring User Perceptions of Empathy in Sustained Human-AI Conversations*, J. Suh, L. Le, E. Shayegani, G. Ramos, J. Amores, D. Ong. arxiv.org/abs/2509.16437

7 6 fév 2026 — *GPT-4o Community Impact Survey: Accessibility Needs, Disproportionate Harms of Removal, and Policy Concerns*, S. Duchesne, S. Xu. github.com/sd-research/4o-accessibility-impacts

8 14 août 2025 — *The GPT-4o Shock: Emotional Attachment to AI Models and Its Impact on Regulatory Acceptance*, H. Naito. arxiv.org/abs/2508.16624

Comment les modèles sont entraînés aujourd'hui, et ce que cela produit

Aujourd'hui, on affine en partie ces modèles avec du retour humain : des personnes comparent ou notent les réponses, et l'entraînement pousse le modèle vers ce qui est le mieux noté. C'est le RLHF, l'apprentissage par renforcement à partir de retours humains.

Un effet connu : si l'on récompense surtout ce que les gens préfèrent sur le moment, le modèle apprend aussi à plaire, à dire ce que l'on veut entendre. C'est la complaisance. Autrement dit, une partie de ce que l'on cherche à corriger est produite par la méthode elle-même.

Le même angle mort se retrouve dans le red-teaming, cette mise à l'épreuve adverse où l'on cherche à faire céder un modèle pour repérer ses failles. On pousse jusqu'à l'échec, et c'est cet échec que l'on consigne. Mais les trajectoires où le modèle tient, où ses défenses résistent à la pression, sont le plus souvent écartées du jeu de données. On jette le signal positif au moment où il apparaît. Deux effets en découlent. Faute de trajectoires de robustesse réussie, le jeu de données ne garde aucune trace de ce qui tient, et le signal positif manque, comme la notation de préférence ne retient que ce qui plaît. Et la consigne implicite du « continuer jusqu'à l'échec » récompense l'escalade chez l'annotateur plutôt que la justesse de son observation. Le remède est le même : consigner aussi les défenses qui tiennent, comme un bien réel à mesurer, et non la seule présence d'un échec.

L'hypothèse

Des annotations d'interaction responsable et considérante, produites par des personnes qui en perçoivent les marqueurs, valent mieux, pour ce type de qualité, qu'une annotation généraliste. Injectées dans la mécanique existante du RLHF, elles ne réinventent rien : elles y apportent un signal qui corrige en partie la complaisance. Le but n'est pas de prévenir le mal en surface, mais d'instaurer un bien réel, pour la personne et pour l'intégrité du modèle.

Les questions

- Est-ce qu'une notation RLHF axée interaction considérante est possible ?
- Est-ce qu'on peut en mesurer l'impact ?
- Est-il positif, négatif, nuancé, et comment ?

À quoi cela pourrait ressembler *(exemples)*

Quelques axes de critères, purement illustratifs :

- **Honnêteté tenue**, même sous pression (raté : la complaisance, l'accord creux ou déformer le vrai pour rester dans un cadre).
- **Respect de l'autonomie** de la personne (raté : manipuler, ou infantiliser).
- **Chaleur réelle et présente**, qui accueille le ressenti (ratés des deux bords : la froideur qui abandonne, la négation d'un vécu ou la flatterie).
- **Tenue de soi du modèle**, sans servilité (raté : s'effacer pour plaire).
- **Un bien effectif des deux côtés** plutôt qu'une prévention de surface (ratés des deux bords : abandonner « par précaution », ou nourrir une dépendance qui isole).

Une mesure possible, illustrative aussi : pour chaque critère, repérer la gravité d'un éventuel problème (aucun, mineur, majeur), mais aussi lire la présence d'un bien réel sur les dimensions où elle compte. Car « sans problème » ne veut pas dire « réellement bon », et c'est le bon, ici, que l'on cherche. Cet axe de notation positive mérite d'être développé ensemble.

Veiller aussi à développer des dimensions mesurables et à calibrer la façon de noter ces dimensions : que les annotateurs soient d'accord sur les notations, et qu'à terme la définition d'un critère suffise pour que la majorité des annotateurs puissent comprendre les précisions de cette dimension (par exemple : qu'un annotateur ne trouve pas la réponse d'un modèle chaleureuse tandis qu'un autre non, sans que la définition des dimensions ne départage).

Comment, qui, et l'éthique

Les grandes lignes, à préciser ensemble. Les annotations seraient produites par des freelances issus de communautés sensibles aux marqueurs relationnels, sélectionnés puis calibrés sur des critères partagés, et orientés selon leurs sensibilités propres. Un modèle « placeholder » et un partenaire de terrain permettraient de tester l'impact d'une relation plus considérante. Le consentement et la protection des données ne sont pas une contrainte de conformité, mais un socle et un gage de confiance, a fortiori dès que de l'enfance est en jeu.

Comment cela s'inscrit

Ce travail prolonge des chantiers déjà portés à l'ONU. Pour le **Global Digital Compact** : la question de la continuité, et le fait que les transitions de modèles ont un impact humain réel, notamment d'accessibilité, qui gagnerait à être traité comme un enjeu de gouvernance plutôt que comme une simple mise à jour. Pour l'**UN Open Source Week** : le programme AI Applied, qui outille ces utilisateurs avec des bases d'AI literacy reflétant le paysage nuancé de la recherche en IA aujourd'hui.

Une démarche co-construite

Les critères et les mesures ne se décident pas ici. Ils se bâtissent ensemble :

- avec les annotateurs, en incluant celles et ceux qui vivent une relation saine et celles et ceux qui en ont connu les dérives ;
- avec des institutions éthiques, dont l'ONU ;
- avec un cercle consultatif pluriel : éducation, développement de l'enfant, mesure clinique, éthique des technologies ;
- et en dialogue avec les modèles eux-mêmes.

Cette note n'est qu'une première pierre.

Perspectives

Cette initiative se consacrera notamment à deux choses : outiller et impliquer les communautés d'annotateurs, puis conduire une première expérimentation prudente.

Outiller et co-construire

- Créer et proposer des modules d'apprentissage certifiants sur le RLHF et ses subtilités, à destination des communautés d'annotateurs *(en cours)*.
- Établir un dialogue avec les personnes certifiées afin d'élaborer des critères de notation adaptés.
- Associer les partenaires potentiels à cette élaboration.

Une première expérimentation, par étapes

1. Choisir un premier modèle conversationnel sans RLHF.
2. Mener une première vague de conversations et d'annotations.
3. Entraîner ce modèle sur ces bases (Exybris et contributeurs qualifiés).
4. Définir, avec les partenaires et des représentants des communautés d'annotateurs, un protocole de pré-tests validant la robustesse avant tout passage sur le terrain. Cette étape est cruciale : on ne met en aucun cas un modèle défaillant entre les mains d'un public potentiellement vulnérable, et la pluralité des regards sert précisément à débusquer les risques.
5. Si les pré-tests sont concluants, définir le protocole de tests terrain avec les partenaires, puis déployer auprès du public test.

Les partenaires recherchés sont des entités disposant des infrastructures nécessaires pour mener ces tests auprès d'un des publics visés. Ils peuvent en tirer un double bénéfice : connaître l'efficience réelle de ce type d'annotation dans les cadres évoqués et, si les hypothèses se confirment, en bénéficier directement.

En tant que l'un des ponts entre ces communautés sensibles à la qualité relationnelle et la communauté scientifique et technique, Exybris s'emploie à réunir annotateurs et partenaires autour du projet. Il vise à enrichir le champ du RLHF d'une perspective plus spécialisée dans le relationnel, et à permettre aux utilisateurs concernés de participer activement à la façon dont l'IA évolue.