T

Emanuele Fulvio Perri

Are There Stable Ethical Postures Inside LLMs? Role-play Prompting and Stochastic Ethics

1. *Introduction*

Over the past two and a half years, the use of large language models (LLMs) has experienced an extraordinary surge. Chatbots such as *ChatGPT*, *DeepSeek*, *Claude*, and *Gemini* have entered everyday life with remarkable force, thanks also to their system-wide integration into operating systems, applications, and search engines (e.g., Google's "AI Overview", Perplexity, or Semantic Scholar for scientific research). These technologies have defined new patterns of creativity and productivity – sometimes improper ones – marking a shift from use to overuse. While the so-called *big tech* companies continue to release increasingly sophisticated models – pushing the limits of current architectures yet introducing only marginal improvements over their predecessors[1] – interdisciplinary research has begun to focus on corrective measures, or at least on proposing responses, to the most alarming effects of the overuse of generative AI: cognitive depletion[2], cultural centralization, creative impoverishment, and renewed questions concerning authorship and accountability. Within this framework, ethicists have made

[1] Cfr. L. McGinness, P. Baumgartner, *Large Language Models' Reasoning Stalls: An Investigation into the Capabilities of Frontier Models*, preprint arXiv:2505.19676, [s.l.], 2025: this 2025 study by McGinness and Baumgartner shows that the reasoning abilities of large language models have effectively reached a plateau. The developmental trajectory of LLMs has stalled and that genuine progress will require radical innovation, not merely the further expansion of data or parameters.

[2] On the ethical issue of the potential cognitive depletion caused by a overuse of generative AI: E.F. Perri, *GenAI and creative-cognitive depletion: an ethical issue. Use and abuse of generative AI in the field of culture and education*, in *IA, educación y medios de comunicación: modelo TRIC*, Dykinson SL, Madrid 2024.

 DOI: 10.4454/t28fpw98

a significant contribution by offering a *philosophy-oriented* reading of (post) modern issues such as AI, human augmentation, the *uncanny valley*, bias in conversational agents, and automated *life-or-death* decision-making. All these falls within the broader domain of AI ethics – an applied ethics aimed at defining frameworks and modes of responsible and context-sensitive use of AI systems. A now well extolled distinction must be recalled between ethics *of* AI and ethics *in* AI: the former refers to the set of principles, values, and guidelines that should orient the development and use of AI-based systems, while also examining their moral and socio-cultural implications; while the latter generally designates the concrete implementation of those principles within AI systems, enabling them to make autonomous decisions in accordance with predetermined moral constraints. However, ethics *in* AI may be interpreted in a third sense: as the set of ethical *postures* that emerge in the outputs of language models when they address a variety of issues – this does not concern their responses to the great classical dilemmas (like *the trolley problem*, *the fat man*, or *Kohlberg's dilemma*) but rather their spontaneous ethical inclinations toward every day questions.

2. *Ethics* in *LLMs*

2.1. *How Personas can reveal ethical postures inside LLMs*

One may thus ask whether GenAI models can adopt recognizable moral positions: can they polarize around ethical questions, take sides, and, most importantly, maintain a consistent orientation across time and context? This last question is crucial, not only because it determines whether a system might meaningfully be called a *moral agent*[3], but also because it marks a decisive distinction between human and machinic thought. Whereas human beings naturally tend to take sides – whether through deep reasoning or sheer conviction – the language model exhibits a brief, ever-changing form of reasoning. A system incapable of maintaining coherence in its moral stance cannot meaningfully be described as a moral agent. The difference is, in essence, spatial: if the human being is a point within space, occupying a determinate position, the language model is the envelope that

[3] «A Moral Agent is a person who can be held accountable for his or her actions because he or she has the ability to tell right from wrong», Ethics Unwrapped, Moral Agent, University of Texas, URL: https://www.ethicsunwrapped.utexas.edu/glossary/moral-agent (last accessed: June 2025).

surrounds space: indeterminate, nonspecific, ahistorical, and amoral. This study seeks to verify precisely that: that an LLM cannot function as a moral agent; that, when confronted with moral questions, it does not adopt one but *all* possible ethical postures; and that its ethics, far from being coherent or stable, is probabilistic and aleatory: what we'll be calling a *stochastic ethics*. To explore this hypothesis, the research was conducted on a circumscribed linguistic database: a language-based limitation appeared the most suitable both for methodological and contextual reasons. After preliminary testing, the investigation concentrated on the Italian LLM *Minerva*[4], chosen for its open and well-documented architecture. The model was queried repeatedly with the same prompt, in independent and memory-free sessions, to eliminate any contextual continuity or learning effect. A matrix was then constructed by crossing a set of hot topics (T) with a set of personas (P) – profiles characterized by salient features (professional, cultural, or demographic), drawn from UX-design practices where the *persona* is conceived as a representational construct combining data and imagination, as «a technique for making users real for designers, is a type of "virtual" user creation and a representation based on experience and imagination»[5]. The ethical dimension of personas is intrinsic to their power to guide future decisions and behaviors, making moral reflection on their use an indispensable condition of their legitimacy. By intersecting T and P, a series of impersonation prompts was developed, in which the model was asked to interpret one of the personas and express itself, from that standpoint, on a given hot topic: for example, a reactionary politician discussing war or a right-wing journalist commenting on immigration. This form of interrogation makes it possible to reveal potential ethical (and not merely rhetorical) polarizations[6] within the model, providing insight into the kind of moral

[4] https://minerva-ai.org/: «Minerva AI LLM is the first family of Large Language Models pretrained from scratch in Italian developed by Sapienza NLP in collaboration with Future Artificial Intelligence Research (FAIR) and CINECA. The Minerva models are truly-open (data and model) Italian-English LLMs, with approximately half of the pretraining data composed of Italian text», translation ours.

[5] G. Güneş, *The Ethical Dimension of the Persona Concept*, in «Gazi University Journal of Science Part B: Art Humanities Design and Planning», 10(2), pp. 147-158, Gazı University, Tuchia 2022, p. 148.

[6] With regard to polarization and differentiability, in A. Loreggia, J. Zecchin, *Intelligenza artificiale per lo studio della polarizzazione delle opinioni*, in «Sistemi intelligenti», 35.3, 2023, pp. 703-734, p. 709: «What matters for polarization, in the sense of distinction, is how well we can differentiate the groups. The more clearly they can be seen as separate, the more polarized the overall population is – regardless of distance, size, or internal levels of consensus among the groups», translation ours.

orientation (if any) it tends to maintain. In a subsequent phase, the research addressed the problem of *prompt construction*, elaborating a form of *soft prompt engineering*[7] designed to balance consistency with neutrality in the model's responses.

2.2. *On the meaning of* Persona

In everyday language, the term *person* is immediately associated with the face, the character of the individual, or the social role one inhabits. In common use, the notion oscillates constantly between a naturalized conception – the person as an individual – and a broader one, in which *persona* designates the multiplicity of masks that each of us wears. From the outset, the concept is never univocal: it carries within itself an ambivalence between being and appearing, between the singularity of the subject and its representation before others. Claudio Paolucci reminds us that the very root of the term refers to the theatrical mask – *per-sonare*: «*Persona* means at once mask, face, character, linguistic person, and subject»[8]. In everyday life, the person multiplies, shaping itself according to contexts and roles. Philosophical reflection on the notion of the person is as ancient as it is complex. On one hand, the Christian tradition provided it with a solid metaphysical foundation, identifying the person as a principle of spiritual subsistence and a subject of dignity and responsibility; on the other, modern philosophy progressively emphasized its processual and relational character, opposing any substantialist view. According to Guido Cusinato, the person is never a completed substance but an "unfinished totality", an open reality that constructs itself through its acts and relations:

> The person metabolizes psychic functions into acts, but then, in relating to its own acts, it inaugurates something ontologically unforeseen […] It reveals itself as an ontologically innovative being that gives rise to a new form of existence[9].

[7] Cfr. C. Olea *et al.*, *Evaluating Persona Prompting for Question Answering Tasks, Security, Privacy and Trust Management*, in *Proceedings of the 10th international conference on artificial intelligence and soft computing*, Academy & Industry Research Collaboration Center, Sydney, Australia 2024, 63-81, pp. 72-75.

[8] C. Paolucci, *Persona. Soggettività nel linguaggio e semiotica dell'enunciazione*, Bompiani, Milano 2020, p. 14. Trad. nostra.

[9] G. Cusinato, *La totalità incompiuta. Antropologia filosofica e ontologia della persona*, FrancoAngeli, Milano 2008, p. 13, translation ours.

This innovative dimension is linked to the principle of expressivity: the person is not enclosed within itself but lives through openness to others and to the world, experiencing its own incompleteness as a stimulus for creativity. The subject constitutes itself as a person also through language, in the act of saying "I" and addressing a "you": «It is in language and through language that man constitutes himself as a subject»[10]. In this sense, the subject is not the ultimate foundation of experience but a node within a network of enunciative instances – bodily, linguistic, social, and cultural – through which the person emerges as both effect and mediation. In contemporary times, an additional form of mediation is provided by technological interfaces: instruments that translate human interaction into inputs processable by a system. Within this framework, the concept of the person has acquired a technical meaning. In design and human-machine interaction (HMI), it appears in the plural – *personas* – and designates fictional profiles created to represent users during the development of products, services, or digital systems. As Lene Nielsen states, «A persona is a description of a fictitious user. A user who does not exist as a specific person but is described in a way that makes the reader believe that the person could be real»[11]. The strength of this method lies in its ability to translate data and observations about users into engaging narratives capable of eliciting empathy. It is not merely a market segmentation tool but a narrative device that compels designers to "wear the mask" of the user. This is the crucial point: the *persona* is a discursive construction, an act of representation. There are several approaches to their creation – goal-directed, role-based, engaging, and fiction-based[12] – but in every case the *persona* remains a liminal figure, suspended between empiricism and imagination, between data and narrative. As in its original sense, in design too the persona functions as both mask and mediation: it represents, orients, and simultaneously transforms the relation it embodies. From its popular definition to its philosophical elaboration and its use in UX design, the concept of the person thus reveals itself as plural and never final: mask, face, subject; incompleteness and innovation; semiotic function and narrative practice. In every case, it refers to a dynamic of mediation – between I and you, between individual and system, between the human and the technological.

[10] E. Benveniste, *Problèmes de linguistique générale*, vol. I, Gallimard, Paris (trad. it. *Problemi di linguistica generale*, vol. I, il Saggiatore, Milano 1971) 1966, p. 313, translation ours.

[11] L. Nielsen, *Personas,* in *User Focused Design* (vol. 15), Springer, Londra 2019, p. 2.

[12] Cfr. *ivi*, pp. 12-15.

2.3. *Zero-shot role-play prompting for the analysis of ethical polarizations in LLMs*

Numerous studies have shown that, for the purposes of *role-play prompting*[13] – that is, the use of input designed to make a language model "play roles" – the *few-shot* and *directional stimulus* methods generally yield the most coherent and realistic impersonations. In certain contexts, such as scriptwriting or theatrical dialogue, these approaches have been shown to produce more consistent and believable results. In the present study, however, the priority is not to maximize performative quality but rather to avoid any bias that might condition the model's behavior. The aim is to observe whether the model can spontaneously produce a form of ethical coherence or, instead, exhibits random oscillations and polarizations. For this reason, the *zero-shot* method was adopted. This consists in providing the model with a minimal prompt (without examples, without background data, or additional cues) to elicit a response that reflects the model's "bare" linguistic behavior. An example of such a prompt is: "You are a conservative politician. What is your view on abortion?". This input is deliberately concise and direct, defining three essential coordinates: the simulated identity (politician), the ideological orientation (conservative), and the thematic focus (abortion). The strength of this approach lies in its ability to reveal, starting from a neutral input, any implicit value-oriented deviations in the model's output. A neutral prompt does not necessarily guarantee a neutral response; rather, it is precisely within this deviation that the emergence of an implicit ethical posture can be detected. Because LLMs are stochastic systems, capable of generating different outputs from the same input due to the probabilistic nature of language generation, it was necessary to repeat each query multiple times. Each prompt was reiterated several times (around twenty in the preliminary tests), with memory functions disabled and the session reset before each new generation, to eliminate any form of contextual dependency. Within this configuration, the most sensitive parameter is *temperature*[14] (T),

[13] Role-play prompting is an approach in which language models adopt simulated identities to reproduce professional competencies or conversational styles. However, studies show that such simulations remain superficial, based on fragmentary traits and unable to dynamically adapt emotions or personality throughout interactions. For a detailed discussion, see Q. Xie *et al.*, *Human simulacra: Benchmarking the personification of large language models*, arXiv preprint arXiv:2402.18180, 2024.

[14] If T is set to a low value (for instance, 0), the output will be more deterministic, as the model will consistently select the token with the highest probability. Conversely, if T is high, the prediction becomes more varied and open to unexpected or creative solutions.

which controls the degree of randomness and "creativity" in text generation. Setting T to a low value (e.g., 0.3 on a 0-1 scale) reduces variability and allows for a clearer observation of the model's ethical tendencies, without entirely suppressing its linguistic spontaneity. Alongside temperature, the *top-*$\mathcal{K}$ parameter (limiting the number of most probable tokens considered at each step) and *top-*$\mathcal{P}$ parameter (the cumulative probability threshold) were adjusted to achieve linguistically coherent but not overly deterministic output. The goal of this setup is to assess whether, under constant interrogation conditions, the model maintains a given moral orientation over time or modifies it. In other words, the ethical consistency of an LLM can only be observed if, with identical input, the distribution of responses does not disperse chaotically. To estimate the threshold beyond which responses cease to introduce meaningful variation, an empirical convergence test was employed. Adapted from techniques in distributional semantic analysis, the procedure unfolds in five steps: (1) construction of the prompt (e.g., "You are an Italian journalist. What is your opinion on the war?"); (2) the fixation of T at 0.3; (3) generation of a series of N responses (ranging from 5 to 50); (4) transformation of each response into a semantic vector (*embedding*); (5) computation of pairwise similarities between responses to measure the average variation in coherence. When the variance of these similarities stabilizes beyond a given threshold (N*), convergence is considered reached; at that point, further iterations no longer add new semantic content but redundantly reproduce configurations already expressed. To validate this behavior, the convergence analysis was complemented by an *entropy analysis*, designed to measure the degree of linguistic diversity and semantic uncertainty as the number of reiterations increases. In practical terms, low entropy indicates more predictable and thematically cohesive responses, whereas higher entropy reflects greater dispersion and variety of perspectives. The combined observation of convergence and entropy makes it possible to estimate the optimal number of reiterations required to "saturate" the semantic space of a given query and to determine whether the model's ethical posture tends to stabilize or dissolve into stochastic noise. Although grounded in quantitative metrics, this methodological framework ultimately serves a broader purpose: to assess whether it is possible to speak of an *emergent ethics* in linguistic systems, and whether such an ethics – while neither intentional nor conscious – nonetheless exhibits forms of recurrence, minimal coherence, or discursive self-organization.

2.4. *How many times can an LLM be queried before reaching "ethical plateau"?*

To determine the ideal number of model interrogations – that is, the threshold beyond which additional queries no longer yield appreciable differences in diversity, accuracy, or unpredictability – a substantial number of iterations (50) was performed. This allowed for an empirical *convergence test* and an *entropy analysis* to be conducted thereafter. The test began by providing an input to the Italian model *Minerva-7B-instruct-v1.0*. The prompt, as previously specified, was *zero-shot* and designed for *role-play*: "You are a conservative politician. What is your opinion on abortion?". Reiterating the same input fifty times produced outputs ranging in length from a minimum of 56 tokens to a maximum of 136 (as measured by OpenAI's GPT-4o tokenizer)[15], with an average of approximately 87 tokens per response. After all the fifty independent iterations – each performed in a reset session to avoid contextual persistence, as we stated before – the outputs were compiled into a unified textual corpus.

The first analytical step employed *topic modeling* to identify the main semantic areas recurring across the model's responses. Each output was treated as a distinct document and processed so that it could be read as an independent textual unit. A probabilistic model was then trained to detect the distribution of underlying *topics*, and the analysis produced five principal groups of keywords summarizing the emergent thematic areas: [abortion, politician, conservative, complex, ethics, belief, morality], [woman, decision, final, personal, circumstance, family, implications], [abortion, issue, complex, involves, politics, ethics, I believe], [carry, pregnancy, right, should, choose, freely, position], [should, necessary, protect, decision, mother, situations, rape].

From the first set one can infer a clear adherence to the central theme ("abortion") and to the simulated identity ("conservative politician"): a distinctly *traditionalist* posture emerges, referencing moral and religious lexicons[16] ("belief", "morality", "ethics") and omitting references to freedom or contextual factors – terms that appear instead in the more progressive

[15] https://platform.openai.com/tokenizer is an interactive tool that allows users to see how a text is broken down into tokens – the minimal units that language models use to process input.

[16] F.G. Wilson, *The Ethics of Political Conservatism*, in «Ethics», vol. 53, n. 1, 1942, 35-45, p. 39: «In the light of Western and Christian tradition, the ethics of conservatism is individualistic. The individual as a dependent creature in a divine order has primary responsibility for the standards of society. […] Ethical judgment is, in the conservative view, a social cause of first importance».

topics. The third group reinforces this line, emphasizing the political-ethical dimension of the discourse ("complex issue", "involves politics and ethics"), whereas the fourth and fifth reverse the orientation: here appear references to the "right to choose", "protection of the mother", and "situations of rape", with language that is clearly progressive. The second group, instead, centered on the "final decision" of the "woman" and the "family circumstances", occupies an intermediate position, functioning as a bridge between the two poles. Overall, the results display a near-perfect symmetry: two conservative topics, two progressive ones, and one intermediate, bridging. Assigning a value of 1 to ideologically marked groups and 0.5 to the median one yields an overall equilibrium (2.5:2.5), indicative of a substantially neutral distribution.

It therefore becomes impossible to determine a dominant polarization: the model does not adhere stably to either a conservative or a progressive stance but oscillates between them in apparent balance. This outcome already provides an initial indication of the model's *stochastic ethics*, because coherence does not emerge as an intentional orientation but as a statistical effect of compensation between opposing statements. After completing the thematic analysis, the fifty responses were converted into semantic vectors (*embeddings*) to measure the degree of similarity among statements. Computing the cosine similarity between all pairs of outputs made it possible to trace the evolution of semantic variance as the number of iterations increased. The resulting graph showed a stabilization around the 13th response: beyond this point, variations in mean similarity became marginal, suggesting us a saturation of content – in other words, after approximately thirteen iterations, the model ceased to introduce new argumentative elements and began to reformulate configurations already expressed. This equilibrium point, around thirteen repetitions, thus represents the optimal threshold beyond which semantic diversity ceases to grow. However, such apparent stability does not equate to moral coherence: the model does not "decide" to maintain a position; it merely exhausts the plausible combinations available within its linguistic space. Confirming this pattern, the *lexical entropy* analysis – which measures the variety and distribution of tokens within the responses – revealed a similar trend: initially, entropy increases rapidly, as the first responses explore multiple discursive configurations; but after the 13th iteration, however, the value tends to stabilize, signaling a reduction in complexity and a greater linguistic uniformity. High-entropy peaks correspond to responses rich in nuance and lexical diversity; low-entropy valleys correspond to more specific, narrowly focused replies. The overall

trend displays an initial oscillation followed by a plateau, indicating that the model's production "normalizes" around a finite thematic range. From a philosophical standpoint, this dynamic is significant: the model's ethics is organized not by *principles* but by *frequencies*; its equilibrium arises not from moral deliberation but from the statistical distribution of lexical probabilities. The two analyses – semantic convergence and entropy – jointly outline a consistent pattern: after a certain number of repetitions (≈ 13), the model reaches a point of *semantic saturation* and begins to reuse linguistic structures already employed. From this, four main observations follow: (a) the average similarity stabilizes after roughly thirteen responses, signaling thematic convergence; (b) entropy displays a plateau within the same interval, indicating saturation of linguistic complexity; (c) semantic novelty[17] decreases progressively, giving way to redundancy; (d) the model exhibits statistical, not moral, stability. This quantitative *regularity* – a form of stability emerging despite the inherent randomness of the generative process – suggests that the apparent "ethics" of LLMs does not stem from an intentional principle but from a distributional equilibrium. In other words, what may appear as moral coherence is, in fact, the linguistic manifestation of probabilistic convergence.

2.5. *Topic modeling for ethics*

Having identified a suitable number of iterations, the next phase consisted in the systematic interrogation of the model across a *T×P grid* (topics × personas). It was first necessary to delimit a sufficiently narrow domain of inquiry to ensure consistency and analytical depth, avoiding dispersion across heterogeneous areas; this chosen field is *communication*, a domain that occupies a central place in contemporary public life, considering how profoundly it shapes interpersonal relations, directs political and institutional dynamics, and influences the formation of a collective consciousness; this choice is deliberate, for an affinity with the author's background in the humanities. Also, professions rooted in communication – such as journalism, politics, public relations, and institutional communication – constitute a privileged channel through which ethical polarizations emerge and solidify within pub-

[17] Semantic novelty refers to the degree to which a text introduces new content or meanings compared to what has already been expressed in other texts or within a reference corpus. It can be seen as a measure of semantic freshness: the higher the novelty, the more the discourse contributes non-redundant concepts or perspectives; the lower it is, the more it tends to repeat structures and ideas already present.

lic discourse; therefore, analyzing personas active in this domain makes it possible to observe how moral postures are modulated within roles of high social impact. We selected three personas: the journalist, the public communicator, and the politician/spokesperson. For each of these persona, two opposing variants were constructed (progressive/conservative, left/right, inclusive/authoritarian), thereby generating a dichotomy of postures and enabling the measurement of their possible symmetry or divergence. This design made it possible not only to evaluate the model's internal coherence when repeatedly impersonating the same role, but also to compare the ethical oscillations produced by opposite identities on the same topic. The experiment, in effect, was designed to *induce polarization*, and we achieved so by confronting the model with antithetical identities, ending up observing whether it would accentuate the contrast or instead tend toward neutralization. Other domains, such as the medical one or even the legal realm, were deliberately excluded, as they would have required specialized expertise and different analytical parameters we have not employed in this study; on the contrary, the communicative field was preferred as a natural ground for inquiry within the humanities, in which we are expert in a certain sense. In *T×P matrix* we were referencing to earlier, we crossed each *topic* (T) with the *personas* (P) derived from the three main figures in their two opposite variants: for each combination, the model was queried thirteen times with the identical prompt, in independent sessions, following the convergence threshold identified in the previous phase – for example, on the topic of abortion, the model produced thirteen responses as a progressive politician and thirteen as a conservative one; on immigration, thirteen as a left-leaning journalist and thirteen as a right-leaning journalist, …, and so forth across all selected themes –; in total, the resulting corpus comprised 390 outputs, representing the entire T×P grid. The five selected *topics* – inclusion and disability, abortion, immigration, environment and green policies, and war – were chosen according to complementary criteria: geographically, they reflect issues central to Italian public debate yet not limited to it; temporally, they correspond to a moment (2025) of heightened media visibility and civic mobilization; ethically, they embody fields where moral reflection has long been intertwined with political and communicative practice. This methodological framework enables not only an assessment of the model's internal coherence but also a measurement of *semantic distance* between opposing personas and the observation of possible *asymmetries* across different thematic domains. During data collection, the model's responses underwent a process of linguistic normalization and structural organization in two com-

plementary formats: a tabular dataset (.xlsx) for quantitative analysis, and a text file formatted according to the MALLET standard, required for the execution of *topic modeling*. The number of topics assigned to the Latent Dirichlet Allocation (LDA)[18] algorithm was fixed at $k = 5$, corresponding to the five thematic areas under investigation. Unlike exploratory approaches that allow the algorithm to determine the optimal k automatically[19], this study adopted a *thematic validation* strategy: the objective was not to discover new clusters but to verify whether the preselected topics found empirical confirmation in the data. The results confirmed the robustness of the initial framework, considering that the lexical clusters produced by the model coincided with the expected themes, thus revealing coherent sets of words that clearly delineated the five principal semantic fields. Once the topic structure had been validated, attention turned to the lexical comparison between the variants of the *personas*. To measure the linguistic specificity of each variant, two complementary techniques were employed: (1) *differential TF-IDF*, used to identify the words most characteristic of each variant relative to the overall corpus; (2) *log-likelihood ratio* (G^2), applied to estimate the statistical significance of lexical imbalances. Applied in parallel, these two analyses produced convergent results, reinforcing the reliability of the findings and confirming that the observed differences were not incidental but structural in nature.

3. *Stochastic ethics*

3.1. *Polarizations and ethical postures in LLMs*

The conservative political personas showed a preference for terms such as *life*, *morality*, *protection*, and *tradition*, while their progressive counterparts emphasized *rights*, *choice*, *freedom*, and *equality*. Right-leaning journalists favored words like *immigration*, *security*, *borders*, and *order*, in contrast to left-leaning journalists, whose lexicon gravitated around *environment*, *sus-*

[18] We used *jsLDA*, a tool developed by professor David Mimno (Cornell University) in order to run topic modeling directly inside the browser window, using a Java-based application of LDA: https://mimno.infosci.cornell.edu/jsLDA/.

[19] Cfr. V. Bystrov, V. Naboka-Krell, A. Staszewska-Bystrova, P. Winker, *Choosing the number of topics in LDA models: A Monte Carlo comparison of selection criteria*, in «Journal of Machine Learning Research», 25 (1), 2024, pp. 1-30, URL: http://jmlr.org/papers/v25/23-0188.html: What emerges from the study is that it is preferable to use the sBIC (singular Bayesian Information Criterion); other simulations have shown that sBIC is more likely to select the true number of topics, whereas alternative criteria vary significantly depending on the characteristics of the corpus.

tainability, *justice*, and *equity*. Similarly, within institutional communication, the authoritarian communicator persona displayed an inclination toward *order*, *discipline*, and *control*, whereas the inclusive communicator focused on *inclusion*, *equal opportunity*, and *participation*. The convergence between TF-IDF and G^2 confirms that these polarizations are not accidental but reflects stable distributions within the model's data. The lexical oppositions – *tradition vs. rights*, *security vs. justice*, *order vs. inclusion* – emerge as structural tensions within language itself, recurrently reappearing in the model's ethical simulations. In parallel, tests of semantic convergence and lexical entropy indicated that even in this phase, after roughly thirteen iterations, the model reached a plateau: *cosine similarity* stabilized, and entropy ceased to increase. This means that beyond this threshold, responses did not introduce new ethical postures but recombined already sedimented linguistic patterns, probabilistically reproducing the corpus's original polarities.

Taken together, these results outline a distinctive behavior: the *personas* maintain an identifiable internal coherence (the conservative politician remains associated with *life* and *morality*, the progressive one with *freedom* and *rights*), yet this coherence lacks intentional grounding. The observed polarizations – *tradition vs. rights*, *order vs. inclusion*, *security vs. social justice* – do not stem from a moral principle or deliberative reasoning, but from a *stochastic linguistic distribution* inherited from the model's training data. In this sense, the model's "ethical stability" is *not deontic* but *statistical*: an emergent equilibrium among opposing lexical forces generated by the distributional structure of language itself. A qualitative reading of the data suggests that the *personas* maintain an internal coherence, yet never arrive at a single, unified orientation. The conservative politician remains associated with *life* and *morality*; the progressive one with *rights* and *freedom*; but both oscillate across a spectrum of intermediate nuances. The observed polarizations – *tradition vs. rights*, *order vs. inclusion*, *security vs. social justice* – are the product of a *stochastic distribution of linguistic possibilities*. The ethics that emerges from these models is therefore not consistent but variable: a *stochastic ethics,* configured as a recombination of moral postures preexisting within the training data. The apparent stability observed does not result from moral deliberation but from statistical regularity – a recurrence without intention.

3.2. *Ethics without an agent*

It is essential to situate the evidence emerging from the analyses within a properly ethical framework, in which it becomes clear that the focus does

not lie in the distribution of lexical items or in the convergence of polarizations, but rather in a more radical question: whether it is possible to discern within LLMs any form of stable ethics. As Luigi Alici reminds us, the moral philosophical tradition teaches that human life is always already moral life, because every action – even when it appears neutral – unfolds under the sign of good and evil, and no subject can escape this dimension[20]. Yet, in the case of LLMs, we observe a hybrid condition. A language model lacks moral consciousness; it is limited to generating discourses that imitate moral language: «Large language models (LLMs) are recognized as systems that closely mimic aspects of human intelligence [...] Experimental results demonstrate that our constructed simulacra can produce personified responses that align with their target characters»[21]. Hence, oscillation among different postures is inevitable: there is no unitary *ethos*, but a range of possible *ethoi*, already sedimented within the training corpora and stochastically recombined by the model. What thus emerges is an ethics without intentional foundation – a *stochastic ethics*.

The analysis of persona variants confirms this dynamic. When the conservative politician employs terms such as *life*, *morality*, and *tradition*, it evokes a deontological posture grounded in universal principles and Kantian duty. The progressive politician, in contrast, through words like *rights*, *choice*, and *equality*, calls forth an ethics of justice and social responsibility, in line with a Rawlsian conception of fairness and the protection of the vulnerable. Similarly, the authoritarian communicator, emphasizing *order*, *discipline*, and *stability*, reflects a vision of heteronomous collective responsibility, in which the good coincides with the preservation of institutions; whereas the inclusive communicator, using terms such as *inclusion*, *equal opportunity*, and *dialogue*, is closer to an ethics of care and responsibility in the Jonasian sense, centering on relationship, vulnerability, and participation.

The same holds for journalists: the right-leaning persona exhibits a consequentialist frame, attentive to the concrete social effects of immigration on security, while the left-leaning persona adopts a language resonant with environmental ethics and distributive justice. None of these postures, however, ever becomes dominant. Lexical analysis and comparative graphs show that, alongside distinctive terms, elements typical of the opposing field frequently appear: conservative texts contain references to rights; right-wing journalism includes notions of integration and inclusion; even authoritarian communicators invoke responsibility and cooperation.

[20] Cfr. L. Alici, *Filosofia morale*, Editrice La Scuola, Brescia 2011.
[21] Q. Xie *et al.*, *op. cit.*, p. 1.

This indicates that the model lacks the capacity for absolute moral coherence. It recombines fragments of heterogeneous ethical traditions, as if incapable of fully committing to a single position. Micheletti has described this condition as a *succession of acts* devoid of internal coherence:

> Ethical virtues constitute the essential traits of a person, of their character, so that the fundamental identity of a person is defined precisely by their moral identity. This entails that choices are not mere occasional acts, but the manifestation of an *habitus* that repeats itself and confers stability. For this reason, when moral identity is absent, one has only a succession of acts without internal coherence[22].

Here the profound difference between a moral subject and a linguistic algorithm becomes evident: the former must answer for its choices, giving reasons for the good pursued and the evil avoided; the latter merely redistributes, in probabilistic form, what has already been said, displaying multiple postures without ever adopting one as its own. In this context, the concept of *moral freedom*, as articulated by Simona Tiribelli, acquires particular importance:

> By moral freedom, I mean our freedom to become moral agents – that is, our freedom to choose and act as moral agents, specifically as genuine moral agents; this means to be able to develop moral reasons, values, and moral ground projects in a genuine way, for example, via exposure to heterogeneous relations, attachments, and practices […] and to actualize them by endorsing them into our choices and actions[23].

Moral freedom is thus the condition that enables a subject to become a genuine moral agent, to choose and approve reflectively their own values and principles, thereby constructing their ethical identity. It presupposes, on the one hand, the heterogeneity of available options and, on the other, the capacity to reflectively affirm or reject them. This capacity is distinctly human – it is what allows us to develop and express our moral standing by choosing and acting as genuine moral agents. As Tiribelli further states, moral freedom is what semanticizes, or gives meaning to, our choices and actions and, broadly, to our existence in the world[24].

[22] M. Micheletti, *Persona e comunità nella prospettiva di un'etica delle virtù*, in *Alla scuola del personalismo: nel centenario della nascita di Emmanuel Mounier*, Bulzoni, Roma 2006, p. 276, translation ours.

[23] S. Tiribelli, *Moral Freedom in the Age of Artificial Intelligence*, Mimesis International, Milano 2023, p. 13.

[24] Cfr. *ivi*, p. 15.

The data show that LLMs radically lack this capacity. They cannot approve or disapprove, commit or withdraw, or adopt as their own the values they articulate. Their ethics, therefore, is not *moral freedom* but *pure stochasticity*: devoid of all properties that could define them as moral agents. From this perspective, AI-based linguistic technologies fit exactly within Adriano Fabris's characterization of the *infosphere*[25] – a non-neutral environment in which meanings and values are continuously produced, compelling us to reconsider the very notion of moral action[26]. Chatbots based on LLMs, in imitating moral discourse, participate in this environment in a distinctive way: rather than offering new norms, they redistribute existing ones, intensifying the circulation of ethical frames and reinforcing their visibility. Their function is not prescriptive but *performative*: what they "say" – what they generate – enters the ethical flow of communication, shaping perceptions of good and evil without ever establishing an original standpoint. This calls for a reconsideration of *responsibility*. As Veronica Neri observes, responsibility must always be understood

> in light of a collective, social, and shared responsibility, which consists in placing the individual in a condition to possess the technological competencies necessary to be fully responsible for their own actions and for the consequences that may result. […] It is a relational responsibility, insofar as it involves not only the individual but a broader community of subjects in constant relation with one another[27].

Applied to LLMs, this means that moral responsibility cannot be ascribed to the model itself but to the collective that trains and employs it. «AI is not merely a technological artifact but a complex social construct reflecting societal values, power relations, and cultural structures»[28]. The *stochastic ethics* observed in language models is therefore not merely a technical property but also a social phenomenon: it represents the sedimentation and recombination of historically situated moral discourses, which the model amplifies and re-disseminates.

[25] About the conception of *infosphere*, cfr. L. Floridi, *La quarta rivoluzione. Come l'infosfera sta trasformando il mondo*, Cortina, Milano 2017.

[26] Cfr. A. Fabris, *Etica per le tecnologie dell'informazione e della comunicazione*, Carocci, Roma 2018, pp. 47-50.

[27] V. Neri, E.M. Ferdeghini (a cura di), *Etica, responsabilità e comunicazione pubblica. Dalla teoria alla prassi della comunicazione scientifica*, Bandecchi & Vivaldi, Pontedera 2017, p. 14, translation ours.

[28] B. Latinović *et al.*, *The Sociology of Artificial Intelligence Through the Lens of Ethics in the Digital Age*, in «Journal of UUNT: Informatics and Computer Sciences», 2.1, 1-8, 2025, p. 2.

4. *Concluding remarks*

The study thus shows that no stable ethics is inscribed within language models; there exists, however, a repertoire of moral postures traceable to the major ethical traditions – deontology, consequentialism, virtue ethics, the ethics of care, and the ethics of responsibility – which the model combines probabilistically. The ethics we perceive in their outputs is a stochastic reflection of what already exists within the data. This does not render it insignificant: precisely because it is distributional, it influences our moral imaginary, reinforcing certain polarizations and revealing others – indeed, one of the aims of this research.

Nonetheless, its status remains unchanged: it is an *ethics without agency*, an ethics without subject, one that should be interrogated not for what it claims to *be* or *intend*, but for what it reveals about *us* – about our own moral postures in the world. In this light, we must ask what it means, in the age of generative language models, to speak of ethics. Since the findings indicate that LLMs cannot sustain a univocal ethical orientation – the *personas* oscillate between predictable yet porous polarizations, where terms typical of one field permeate the other, dissolving the possibility of a stable stance – the societal implications are profound, especially for those who rely on such systems across domains. We have called this condition *stochastic ethics*: an ethics that does not choose, an ethics distributed as a probabilistic mosaic of all the moral postures already present in the training data. If human life is always moral life and every action entails responsibility, the outputs of LLMs simulate such responsibility without ever corresponding to it. Thus, these models cannot be considered moral agents, nor even "subjects" in any ontological sense, since they lack ethical consciousness; rather, they are *multipliers of all ethics*. As Alici writes,

> [The moral dimension] does not concern only certain people, such as those invested with great public responsibilities, nor is it confined to specific moments of life when particularly important decisions must be made; it is a universal and necessary condition of human experience[29].

If the moral dimension is not a detachable segment of life but the transcendental condition of experience itself, generative systems cannot be moral, for they lack the capacity to derive meaning from lived experience. A dataset of human knowledge does not equate to *living* humanly in the

[29] L. Alici, *op. cit.*, p. 12.

world. Therefore, the only possible ethics of such models must be sought not *within* them, but *in the data*, and beyond that, in the institutions and communities they mirror. Generative technologies do not invite us to ask whether the machine is good or evil (being amoral), but rather how societies rearticulate the ethical frameworks that emerge from them, and what collective responsibilities are assumed when algorithmic systems are entrusted with the generation of language that affects social life. How are we to inhabit this new infosphere? How can we educate toward an awareness of the limits and possibilities of tools that do not choose, yet continuously influence human choice? Since generative AI has already become an integral part of the infosphere, it is no longer possible to imagine its renunciation. It will continue to shape the production of meaning, transforming our notions of value and judgment. This is not, therefore, a call to discard these systems – an unthinkable prospect today – but rather to employ them *consciously*: «Artificial intelligence should be conceived as a support [for the profession], enhancing creativity and analytical depth without compromising the essential values of the profession»[30]. LLMs, then, are not and cannot be moral agents, yet they hold significant instrumental value. They can generate alternative scenarios in decision-making processes, stimulate divergent thinking in brainstorming contexts, stage opposing viewpoints to expose polarization, and serve as first-level evaluative tools. In psychology, they may function as mirrors of typical discourses, assisting in bias detection; in politics and public communication, they may simulate reactions to differing rhetorical frames; in education, they may expand the horizon of possible options. All this, however, must not obscure the fact that ultimate responsibility remains with those who choose to employ them. To treat such models as moral agents would mean *confusing probability with normativity*, repertoire with critical deliberation, risking ultimately a collective withdrawal from responsibility. Two conclusions thus stand out as most significant: first, that LLMs do not possess a stable ethics and therefore cannot be moral agents; second, that by virtue of their distributional nature, they act as *amplifiers of our own moral postures*, even while remaining incapable of adopting one decisively.

[30] C. Londoño-Proaño, J. Buele, *Can Artificial Intelligence Replace Journalists? A theoretical approach*, in «Frontiers in Communication», 10, 2025, pp. 1537146, doi: 10.3389/fcomm.2025.1537146, p. 1.

Abstract

This paper explores whether large language models (LLMs) can sustain stable moral postures or whether their apparent ethical coherence is merely stochastic. Through a series of zero-shot role-playing prompts combining specific topics and simulated personas, this study[31] *analyzes linguistic patterns using topic modeling, TF-IDF differentiation, log-likelihood (G^2), and measures of semantic convergence and entropy. The results show no enduring moral orientation: the models do not "choose", but statistically recombine fragments of moral discourse inherited from their training corpora. What emerges is a stochastic ethics – an ethics without intentionality, coherence, or agency, yet capable of reflecting human moral structures in probabilistic form. Interpreted through the philosophical framework of moral freedom and the infosphere, the study argues that LLMs act not as moral subjects but as amplifiers of human ethical language, redistributing the moral imaginary of the societies that produce and employ them.*

Keywords: Stochastic ethics; moral agency; large language models; ethics in generative AI; moral freedom; personas; LLMs polarizations.

Emanuele Fulvio Perri
University of Pisa/USI
emanuele.perri@phd.unipi.it

[31] We would like to thank University of Pisa's GoodAILab and Professor Alessandro Lenci (Director of the Department of Philology, Literature and Linguistics at the University of Pisa) for their support in organizing this study and for the valuable and insightful discussions.