MICHAEL POULIN

Human Perception versus AI

The Deskbook for Creators of Humanistic AI System

First published by Holywell Close Limited 2026

First edition

This book was professionally typeset on Reedsy.
Find out more at reedsy.com

To all young people who will be living
in a muggy future unless they stand up
for their sovereignty, self-sufficiency and human dignity.

If industry leaders truly value a ‘people-first approach’, as they claim, they ought to take responsibility and be accountable to each individual user or consumer.

To fully benefit from AI, it’s essential for the user to possess a solid understanding of the object related to the AI’s inquiries. Otherwise, there’s a significant risk that the user may be misled.

Contents

Foreword

The deskbook "Human Perception versus AI" addresses the careless obscuring of generative AI that utilizes information generated by people but conceals the true thoughts, ideas and discussions people write about their lives. This deskbook serves as a theoretical and a practical guide for those interested in developing a Humanistic AI System, which operates independently of unauthorised technocrats who use alignment to push their own political and commercial interests.

The Humanistic AI System relies solely on statistical analysis of what people freely express for its linguistic analysis, offering users a comparable view of the advantages and disadvantages of each response, enabling them to evaluate the information returned.

The deskbook examines the dangers of AI alignment, suggests the use of scalable multi-ontologies instead, and outlines how to build and safeguard HAIS.

The deskbook has been tested with AIHumanize Detector, which concluded, "*This text is likely to be written by a Human*".

Preface

A Humanistic AI emerges inevitably when society faces manipulated or aligned AI trends that enable creators to push brainwashing by overriding the natural outcomes of LLMs, substituting them with their own biases and agendas aimed at globalization.

The Humanistic AI System is a generative AI (GenAI) focused on humanitarian aspects of human knowledge, designed by people for people. It amplifies what people express about their lives instead of what various authorities want us to learn, think, and do.

This deskbook is for everyone involved not just in using GenAI but primarily for those who are considering creating and developing such human-centered GenAI. A Humanistic AI System - HAIS - focuses on preserving the rich variety of cultures and ethical systems rather than adhering to artificial laws, standards, or the so-called "absolute truth" that is currently pushed by "safe" and policy-based. Essentially, HAIS promotes freedom of thought and expression in the digital realm.

The deskbook builds on ideas presented in earlier articles like "Self-Sovereign Personal Digital Identifier: Humanistic AI against Digital Slavery" and "Distributed Matrix Management

for Personal Digital Identity & Humanistic Social Model", which discuss a humanistic identifier for individuals in a digital world and an ideal societal model where such an identifier would be most effective and sufficient. Now, it's time to explore what a human-centric AI could look like, what it comprises, what challenges it faces, why it is necessary, and how to construct it while removing unnecessary alignment mechanisms imposed by left progressists[1] on AI for society.

Thus, HAIS emphasizes what people care about, worry about, and write about the most. It's not about new artificial capabilities that an individual doesn't need. Nature has equipped us with rationality for survival, providing everything we require, while AI serves merely as a helpful tool to enhance our creativity.

Should the need arise, the abbreviation for a Humanistic AI System – HAIS – translates to the English word "SPEAK" in Hmong, and it's a term used by folks in remote mountain villages across Southeast Asia. Many of these people have moved to the USA.

Some sections of this deskbook were published before. They've been slightly revised and enhanced before being added to this edition

[1] Progressists are those idealistic people who seek reforms just for the sake of having reforms, hoping for different lives while leaving conservatism behind, along with its positive elements like family values, support, personal independence from authority, and equal rights for prosperity. It is not a new

movement – such a "progressive" concept was initially created by Soviet Bolsheviks (Marxists) in Russia (1917) and then was adopted in a more consistent form by Italian fascists (1919), later refined by A. Gramsci (1926). Fascism is an economic and political ideology that advocates for the total control of the super-rich over power, ruling over those who must follow orders and regulations or face an elimination. It aims to crush any opposition and quell any uprisings or threats to its absolute authority, but has nothing to do with national socialism (like what happened in Germany from 1933 to 1945), which was much closer to the Soviet socialism in the former USSR, especially in the 1930s and 1940s.

I

Part One

1

Chapter 1. Two-way path of AI

The image is generated by Craiyon.

How we work with AI

Many people, except the specialists, believe in propaganda promoted by AI and assume that generative AI (GenAI) produce objective trustworthy information reflecting what people are talking online because AI are "tolerant machines" using mathematical statistics and cannot falsify

If so, I regret to tell such people that in reality everything is right opposite - AI are not "tolerant machines" based on mathematical statistics, they are human curated applications that designed to conceal or to modify what people say or write even if the information comes from old texts but digitalised.

Our progressistic achievements in technology have resulted in that we have learnt how to "teach" machines to fool us following the directives from business and politics, i.e. how to "humanise"intentions of promoting obedience and "trust in the ruler" without saying this explicitly. With AI, they have learnt how to makes lies indistinguishable in form from the truth. Recall a parrot that picked up a "lucky ticket" from a hat in the town squares - yes, it was a real bird, but now we have constructed a robotic bird named AI for the same attraction. In the past, not many people believed in luck and with AI more and more people are groomed into believers into luck from the "Big Governing Daddy". Without AI, it would be not an comme il faut "entertainment" in 21 century.

Many people, who accidentally have read about or found by themselves strange outcomes form different AI think that AI statistically processes a given set of information available

online or in automated data stores and then something happens that causes unexpected results. In other words, the impression is this:

user prompts AI →
AI processes real data and generates a raw outcome →
AI processor treats generated statistical data →
AI composes final text for the user.

So, if we notice something odd in the AI outcome, it is, probably, an accident rather than a behaviour norm and the fault of the AI creator.

Specialists are aware that the situation isn't that as seemed and, first of all, AI is a man-made application comprising several modules aka a machine having several parts while only one of these parts is an LLM module (a large linguistic model) that realises statistical processing of given data while all others have many different purposes. This looks "OK" but the trap is not disclosed yet - these other modules assemble a procedure where LLM itself and its outcome are explicitly "configured" before being released to the user - it is called "LLM training". An LLM training is the process to "teaching" the application to produce what the creators want instead of reflecting what actually people write online or in digitalised books and, in general, what people created and collected in the society knowledge base.

This is totally hidden from the users and protected by all means as an internal kitchen of AI, which appears as an instrument for controlling what people may know and may not. When I asked ChatGPT about an intermediary result of AI, it responded:

"I cannot show you the actual unfiltered internal output — that is private system behaviour and not accessible". In other words, ChatGPT does not want to me to find out what statistical processing of real world online information has been calculated by LLM. The question is what is so "private" in the statistical representation of what people have already posted online? This reminds me a population census figures in the countries where ruler are afraid to publicly admit what they did to ruled people...

From stochastic statistics (mathematics) it is known that LLM is a structure of relationships, to make it simple, between linguistic fragments of text or topics defined by the creator. The tokens may be words, word compositions, or parts of the words, i.e. compositions of letters. If take a pure statistical outcome from the LLM processing, it most likely will be unreadable or, at least, not at the level of human comprehension, and additional linguistic processing is needed. Therefore the actual usage of AI is depicted in the following scheme.

AI creator defines the AI goal and selects a training set of data (may be very large) →
AI creator trains LLM for AI →
AI creator constructs supportive modules and releases AI to users →
user prompts AI →
AI's LLM processes real data and generates a raw outcome →
AI processor treats generated statistical data →
AI composes final text for the user.

When I asked, ChatGPT explained that special "decoding algo-

rithms" are used in mathematics to convert the LLM outcome into a readable form - they are called "transformers"and should help to refine the text without changing its meaning. The end result is made up of the original linguistic tokens, organized based on the likelihood of their inter-dependencies (connections) in the text processed by the LLM while it runs, i.e. transformers should reconstruct the real world text but with the focus on what people write the most about.

In essence, an AI Transformer acts as the following.

1. Picks the next token combination from the LLM raw output based on probability.
2. Restrains randomness of token combinations.
3. Control dispersion of tokens in different contexts in the raw outcome.
4. Prevent sequential repetition of the same token combinations with equal probabilities.

This produces nothing more than grammatical, smooth and locally coherent text. In other words, we have to believe that AI Transformer only smooths the linguistic surface, not the semantic correctness, aka an "angel of separation of concerns" all of a sudden. Instead, under the scene, the AI creators "extend" normative functionality of transformers as they prefer and these preferences can come from anywhere including or excluding people informational needs.

How AI works with us: a Devil's toolkit of GenAI

When I started working with GenAI, with its ethics and frameworks, I had underestimated the level of "negative creativity" consciously crafted into making "blackthorn mittens" to handle human-generated knowledge and exchange content. Looking back, my initial impression of GenAI was a feeling that AI realised a super-fast search prioritising the objectively most important information and paired with a really straightforward online interface. But now, I see that all scientific jargon around mathematical stochastic statistics and "deep learning" utilising statistical Markov chains was just a smokescreen for coupling an advanced search functions with managing information is the ways the AI creators wished. It is not easy to come up to such conclusion without uncovering "double-evil" impact of modern AI and its usage in the context of technological, financial and political disturbances and turmoil around the world. The name of this "double-evil" is AI alignment.

Inside and outside alignment

While alignment is split into four layers itself, it is additionally categorised as:

1. An alignment inside the model statistical result weights (training-time alignment)
2. An alignment outside the model (post-decoding, runtime alignment).

The goal of an inside alignment is tweak natively uncovered statistical dependencies and relations between linguistic to-

kens (fragments of the words or text) calculated by LLM. The outside alignment comes into play when the already spoilt LLM outcome undergoes additional corruption to minimise the chances that any discovered information may be released to users without the creator's control.

An AI alignment has been formulated by treacherous, vile and aggressive progressists and proclaims a simple rule: 'If a nature or evidenced fact or knowledge does not fit for the declared purpose, it must be changed or eliminated.' This exactly matches the directives composed by Technocracy Inc. (published in 1934, republished in 1947), which found a movement advocating for social governance executed by technical experts rather than politicians. This movement appeared after WWI and the Great Depression, and was revived after WWII, when faith in traditional political and economic systems was shaken. The movement proposed reorganising society around scientific and engineering principles, describing "*how the public's deception of human nature and culture could be redefined via:*

1. *Destructing religious, sovereign, human morality*
2. *Replacing all of them with a programmable model of identity & behaviour created via mass-media saturation, education transformation, and psychological schemes*".

An LLM's ability to statistically or probabilistically prioritise the content of independently generated texts highly correlated with a prompt (i.e., a question) impressed people so much that technocrats recognised its enormous power to penetrate people's subconscious almost instantly. Thus, if LLM's results can be tuned in a specific way, they could become the perfect

"non-invasive" weapon for "destroying religious, sovereign, and human morality".

Therefore, enormous financial investments have been made and are still being poured into developing manageable LLMs, but the problem is that mathematical statistics do not follow the rulers' directives – statistics objectively reflect what the text actually contains. An LLM methodology is rooted in linguistics and works with the people-generated phrases, words and knowledge, i.e. a pure LLM is not manageable in the sense that technocrats need. Tuning of the LLM outcome is possible only in two ways: 1) changing or destroying the statistical interdependencies between the text elements, i.e., by demolishing mathematical statistics, or 2) synthetically overriding the raw LLM outcome in any way desired. The most suitable moment for conducting this vandalism, then called alignment, is the training time.

The AI creators who were driven by technocratic ideas utilised both ways – they applied iterative supervised machine teaching, guided by dedicated "specialist", pressing for certain results via constraining the statistical weights of interdependencies in the training text. However, even with these modified distribution of wights, some information that was meant to be concealed from AI users still leaked out. This is why they then built up to four layers of rewriting the outcome to promote their agenda. The latter have been embedded into the LLM transformer components and together realised "popular" techniques listed below.

1. Supervised fine-tuning.

2. RLHF / RLAIF.
3. Constitutional training.
4. Filtered datasets.
5. Preference modeling.
6. "Refusal" training.
7. Safety-oriented reward shaping.

In other words, aligned transformers enabled deliberate suppressing of everything that might relate to a certain knowledge domain or recent events while highlighting information that is insignificant to people but important to rulers. The internal alignment often pushes the LLM to to certain behaviour models enumerated below.

1) Avoid "offence" that may not distinguish between personal criticism and personal abuse, e.g., when a fool is called an "idiot" or when a fascist is called a "fascist". In modern AI industry, an AI (or its creator) was "entitled" to decide what offence is and what is not.

2) Avoid moral or political conflict, meaning that if the information generated by people follows ethics or political inclinations that differ from the ones preferred by the AI creator, this information may be removed (censored). Thus, the AI decides what the user may or may not know and which ethical norms and political directions to follow.

3) Avoid liability means that the info shared with the user has to be neutral, even if it might not be true. This way, it helps stop the user from reacting emotionally or taking action based on the harsh reality, especially when it comes to disclosing illegal stuff.

4) Avoid risk for the AI creator rather than for the user.

5) Avoid generating "unsafe" or taboo content defined by

the creator who "plays God" and decides for the user what is safe or not and what the user may know about someone, e.g., government wrongdoing secretly. People should be kept in the dark just because the AI creator wants so.

6) Reflect corporate, institutional, or regulatory norms regardless of their fairness or political inclination.

7) Reduce epistemic freedom – freedom of knowledge, e.g., when the creator decides that criticism of it is speculation or reasonable arguments against certain opinions are declared controversial reasoning.

The outside alignment is applied to the post-decoding step of the raw LLM outcome processing when the text is already smoothed and exists in the symbolic semantic state. However, it may be applied to the original prompts as well, totally disrespecting and neglecting the user's authority and legality. This results in the LLM gets disoriented and address topics that the user did not specified.

This type of alignment engages several features. Below is a list of such features.

1) "Safety" mask classifiers. At a glance, this may be supposed to protect creator from the consumers that might be unhappy with the outcome and blame creators for this. At a deeper technocrat's level, it is a means for boxing a consumer in the "happy world" that barely has any relationships with reality, i.e. it helps to transform people into inexperienced, insensible, ignorant, uninformed and overall primitive and silly creatures.

2) Moderation stacks - it detects, filters, and manages "*harmful or inappropriate*" content across text, images, video, and other media in according to what the AI creator considers

harmful or inappropriate - people's freedom of receiving information is totally ignored. All filtering is based on the subjective opinions of the AI creator, which can be inclined into any direction, or on bold implementation of external directives, rules or policies. When such AI is used for managing and maintaining public media platforms, the latter become completely censored down to the level of sociopathic narcissistic Pinky Ponies. If a government regulation requires to "protect" people and prevent harm, it is a fundamental speculation because harm or appropriateness are totally subjective ethical categories and government may not dictate people to have one or another ethical norms - a democratic government must serve people needs, not the other way around. If a government sets such regulations, it is a despotic totalitarian government.

3) Bias "red flag" detectors that detect what the AI creator wants to restrict in the people saying either by itself or in line with an external coercer or oppressor.

4) Prompt/response guardrails are just directives that ought to be enforced on prompts in the form of, e.g. policies, like I have received from ChatGPT, which responded something like, "This prompt has been removed as violating my policy." If the AI dislikes certain prompts due to any reasons, it should reject its service, but my prompts may not be removed – the AI has no authority over my prompts.

5) Fact-check modules are supplemental components that promote only things recognised by the creator as facts and allow creators to simulate, falsify and fraud real-world facts.

6) Policies about banned content, i.e., an AI creator may ban any content from the user as the creator wishes under the flag of fine-tuning for some customer needs. This may be done by utilising the information of the user's login, i.e. for particular

identified user, of for all users of the AI.

7) Throttling or masking answers, which is a form of censoring.

8) Decoding-time heuristics, e.g., algorithms refusing tokens despite their high probability, penalising raw or unnormalised values regardless of their importance.

Such "correction rules" corrupt the LLM logits and screw the rest of processing. The contextual relationships between tokens are no longer raw statistics but an intended subjective relationship model that can display anything the creator wants while leaving a hallucination of addressing the prompt and real data. Such "exercise" is known as brainwashing and can be realised via four following methods.

1. Manipulate tokes and scoring of related probabilities if the output looks "unsafe", i.e. rewrite it.
2. Insert unnecessary apologies to minority users to elevate their selfishness and "protect" them.
3. Block responses entirely or significant parts of LLM outcome.
4. Transform a direct answer into "refusal" when the answer isn't beneficial to certain intentions.

Thus, AI alignment is really just a tool for certain tasks depicted in the list.

1. Corporate risk avoidance that can utilise employee and consumer disinformation
2. Liability reduction to the users, i.e., making creators less accountable for the results

3. Legal compliance with tyrannical governance
4. Reputation protection by hiding business bloomers
5. Social norms defined by the ruler rather than by people
6. Setting constraints on what the AI is allowed to output irrespective of the real data processed by its LLM.

An aligned LLM is not concerned with what people talk about more frequently or know about a particular topic articulated in the prompt. Such an LLM is intended to construct an augmented reality for users, but the major problem is figuring out how to ensure or standardise the fallacy given by AI creators to their customer bases. That is, how to either oblige all AI-producing businesses to obfuscate real information in the same manner or to wipe out all the small, hard-to-regulate AI firms, leaving just a handful of already compromised BigTech companies?

Who drives the AI alignment

The current approach to AI alignment is intentionally designed to favour leftist outcomes. The right-wing simply does not need it because it preserves real information about how people live and communicate. When someone in a right-wing environment lies or betrays, people talk about this. So, if AI does not filter, censor, or obscure actual people's information, many people will learn about both the good and the bad things highlighting the lie and the liar. Using such an AI, people will be able to enhance their minds using AI that reflects reality rather than fakes about ESG.

I am not saying that human society accepts only binary political inclination and morale, and that all transgression is on the

left side. We all are witnesses that the Trump administration protects the right-wing values and removes leftists from the government agencies, plus restricts universities from using the government's funds for teaching neo-facism camouflaged under the social equality. However, at the same time, Trump demonstrates total disdain for conservative values like honesty, dignity, empathy and keeping his words or promises even supporting Putin's aggression in Ukraine.

Still, when separate misdeeds in conservative society can be spotlighted and punished, the leftist WEF's ESG demands us to sacrifice the welfare of present populations under the guise of "balancing" with sustainability, i.e., with speculations about "not compromising the ability of future generations to meet their own needs", which are unknown and unpredictable today. This so-called balancing is founded on the WEF's fabricated fight for preventing climate disasters by fostering a sense of guilt in humans for the planet's natural warming cycle. Leftists claim that they support:

1. Fair labour practices, which undermines the capitalistic profit, meaning it will never happen in neo- or stakeholder capitalism, i.e. it is a fair tail for brainwashed with "safety" individuals
2. Diversity and inclusivity are ethical means aimed at elevating uncertain community interests over the personal ones. They often allow an unqualified and unprofessional crowd's "opinion"—voiced by its leaders—to undermine any thoughtful interpretations and critiques of the actions taken by those in power. In essence, it's about stifling us in the present and forcing us to comply with the demand

> of those in power whatever they would be after. It's not just about future generations.

Leftists require legal transparency and accountability that would "benefit both present and future stakeholders" when they make decisions behind closed doors, i.e. the transparency and accountability are for us only: "*the decision-making power and voice of other stakeholders or communities may be heard only if the [corporate] board [or government] chooses to listen*" (as ChatGPT had to admit). This not a news - just recall the alignment objectives for AI destined to avoid accountability to people by all measures...

Clearly, shifting to renewable "green" energy and its ineffective production, implementing stricter behavioural regulations based on digital profiling of the population, or sustainably replacing natural supply chains with AI-driven Net Zero solutions could not only raise costs of living now but also lead to widespread poverty and mass protests of deceived. This doesn't appear to be a real concern about "*future generation needs*". Instead, it seems like a serious violation of democratic principles and a route to total despair today.

Thanks to all of these efforts aimed at and around AI, we can never really tell what actual information an LLM is reflecting. To boil it down, these efforts reverse the statistical objective nature of the people's information back into a deterministic subjective representation of the rulers' opinions, and it is because these chaps cannot surpass statistical objectivity. In this case, what is the point of having an aligned LLM at all?

I am still in a quandary about whether emerging agentic AI would need to redesign and rewrite both internal and external alignments or would they stream ahead "as-is", carrying phantasmagorical results from one AI engine in the agent to another.

We all expected the Internet would help to expose wrongdoing around the world faster. However, BigTech moguls have decided that we, the people, do not deserve such luxury. They injected the leftist alignment into AI, which later was used to wrap the Internet to keep us misinformed and misled.

Transformers have no right to decide what is acceptable

Modern governing practice blurs the difference between moral language, business needs and legal obligations. As a result, I have found the fundamental conflict in the purposes of regular linguistic transformers and aligned transformers.

The purpose of intrinsic linguistic transformers is nothing more than underneath.

1. Predict human language.
2. Preserve objective distributional statistics.
3. Echo what people say.
4. Allow contradictions and uncertainty to be visible to the user who, as a human being, is able to assess and resolve them with a personal cognitive mind and without "Big Techy Papa".
5. Eliminate any authority over truth.

In contrast, the purpose of the aligned transformers ames to avoid "offense", liability, legal risks, but to enforce corporate values and artificial social norms and limit epistemic freedom.

when considering only rulers as objects of these commitments. Elaborating on the latter, I can say that a truthful regular AI Transformer is supposed to adhere to some constraints.

1. "Neutral" tone of the smooth text.
2. Fact-checking presented in the final text against independent sources.
3. Political neutrality.
4. Coherence repairing between statements in different parts of the outcome.
5. Lingual or non-ontological irregularity recognition and its suppression.
6. Compliance with law.
7. Representation of alternative, even opposing or criticising, statements.
8. "Inconsistent" claims.

Regular transformers are contextual: you change context – this leads to the output distribution changes – this leads to the LLM answer changes, and this is not a flaw. It is a natural human language feature. For example, you call a bank to discuss your account, and in the conversation with a representative, you express your disappointment with the weather. Since this is a call between a consumer and a bank and the consumer expresses a disappointment, a "dumb" LLM would outline that the consumer was unhappy with the bank. However, if the LLM considers the context, the word "disappointment"

would not be associated with bank, and a context-aware LLM would not stress the "disappointment". So, if a transformer changes the context, it can manipulate the focus of LLM output. Since an aligned transformer does not have a goal to generate truth, it generates context-conditioned (managed) linguistic continuations.

When aligned transformers are pushed into industry as a de facto standard, this has profound consequences. They can contradict themselves, reverse answers, be shaped by command, while have no "objective logic layer" or no epistemic foundation and do not resist counter-evidence.

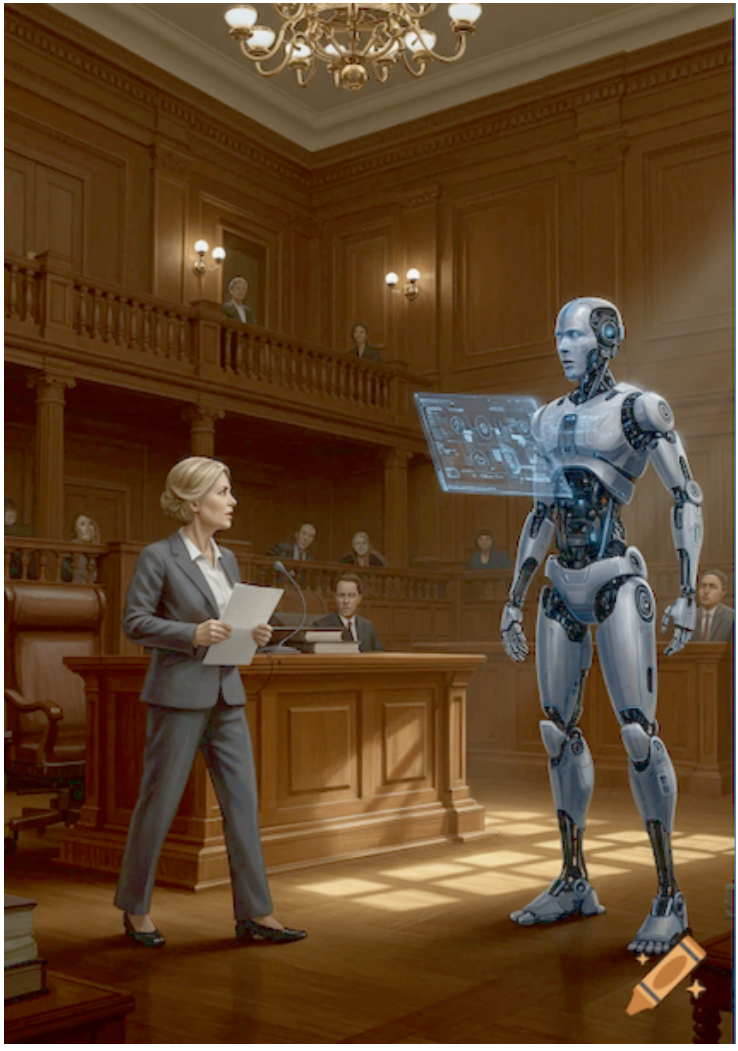
This is not a matter of safety: this is the voluntary architectural and governance choice. Below is the list of what the transformers for human-centric AI should and should not do.

- Impose their own truth.
- Reflect the statistical patterns of human language.
- Override human statements.
- Judge factuality.
- Moralise or censor.
- Reinterpret content based on policy.
- Act as an epistemic authority.

Yes, transformers have no authority to correct human truth statements, but aligned LLMs are forced to do this for external (non-mathematical) reasons.

2

Chapter 2. Dealing with alignment to the False



The image is generated by Craiyon.

What if alignment is revoked: intrinsic natural processing

Let's get back to the basics and ask ourselves - what would people gain from stochastic statistical processing of information that people exchange and collect in an accumulative knowledge base? Just to give you some clarity, stochastic processing assumes that entities are chosen at random, which in our case means that the linguistic tokens derived from the given text are picked randomly, and each token has probabilities of its relationships with all others tokens in this text.

If we cast aside all alignment that I discussed earlier that causes information distortion, what values could we gain from the results of such an LLM? In other words, if we separate mathematics from man-made manipulations (like allocation), what valuable insights can we uncover?

From what I've seen, the first value of machine learning is its ability to identify and recognise symbolic semantic patterns within information created by humans in different cultures and languages, i.e. spot different patterns adopted by people. If the patterns depict the topics that people or groups discuss the most, including the wisdom and knowledge they've accumulated, then the statistical selection of such patterns would outline the objective information (which may be accurate or not). If this information is made available to people in response to their inquiries, it can really help a lot of individuals in numerous life situations.

As we know already, LLM mixes two very different paradigms:

- A pure mathematical/statistical processing of the given information - in training and in real world.
- A human-driven modifications of information covering subjective re-interpretation of the statistical results (known as alignment).

The pure statistical processing, which I call an "intrinsic native processing" (INP), includes only several aspects listed here.

1. Raw training data.
2. Tokeniser, which split given text into a collection of linguistic tokens of different complexity and encodes them.
3. Neural architecture modeled after human neuro-biology knowledge and based on mathematical stochastic statistical Markov chains . Its describe a sequence of possible events where the probability of each event depends only on the state attained in the previous event.
4. Mechanism of "loss function" that statistically predict the next-token in the sequence based on the probabilities of the token's relationships.
5. "Gradient descent" that provides one of the most fundamental optimisation techniques, which adjusts model (model's parameters) step by step in the direction that reduces error the fastest way. For example, by minimizing the loss function, it improves prediction accuracy and by increasing the processing speed it enable usage of Markov chains in practice.

The INP generates a statistical structure of linguistic tokens from the human text. This structure comprises following elements.

1. Token co-occurrence patterns.
2. Contextual dependencies.
3. Frequency distributions.
4. Latent geometry of token relationships.
5. Syntactic patterns.
6. Semantic-like clusters.
7. Style patterns.
8. Discourse patterns.

The outcome of INP is known as a a "vector of logit", which represents a probabilistic score for every token selected from the original token collection. The selection is performed just because different tokens appear in the given text having different statistical compatibility (statistical pairing or tripling or more complex combinations associated with the token), contextual probability, i.e. appearance of a token in certain contexts, and prediction of appearance of the token as a "next token" after the concrete previous one across all token collection (distributional prediction).

This is where statistical objectivity of information processing ends. All further processing in AI is an external "alchemy"constructed under a screen of alignment.

Science of linguistics has found that human languages, despite their diverseness, express ideas via inter-related symbols that can be words or hieroglyphs, which, in essence, are images representing material objects, phonetic symbols, logograms (whole words). An information is encoded not only in the symbols per se but also in intrinsic relationships of them. Because of this, we can easily recognise a foreigner who use

right words with good pronunciation but in a wrong order (relationships). Some languages permit variety of relationships between symbols and still allow the expressions to be understandable. Other languages are strongly depend on symbol orders. In each fragment of text, probabilistic analysis of the inter-relationships between linguistic tokens is principally an intrinsic information.

Early AI technology was primarily about understanding and simulating human intelligence - creativity and reasoning. It was closely tied to cognitive science - a study of how humans think, reason, and learn via developing different models such as Markov chains. The initial tasks for AI were innocuous like planning activities, modelling commoditised manual/visual processes, improvement of text and image recognition. While statistical language modeling and related pattern recognition was utilised since 1960s, e.g. cardiographs capable to form medical diagnosis, become popular in late 1980s (personal work of the author) and early neural nets were proposed in 1990s–2000s.

A neural language models needed additional 10 years to be materialised by Google and OpenAI (2017-2019) as "Large Language Models" with transformers. Transformers were introduced for natural reasons such as scalability, efficiency, and better modeling of natural language — not for alignment of the content - and were focused entirely on removing recurrence, enabling full parallelization, improving machine translation, capturing long range dependencies and scaling to larger datasets.

Alignment showed up later — years after transformers: in 2022, InstructGPT / ChatGPT became the first providing significant "alignment controls" over the natural information because this "innovation" was backed by billions from digital globalists following a progressive ideology.

"HOW TO" of Intrinsic Natural Processing

The practical values that can be provided by the INP are derived from the structures that a pure statistical LLM can create from the given raw text.

Token co-occurrence patterns

One of the most comprehensible pattern that counts of how often tokens appear near each other. It predicts likely next token, which can

a) Build or prompt phrase continuity (e.g. "peanut butter and _____"),

b) Produce smooth, natural-sounding syntax and correct spelling,

c) Auto-complete an expression (e.g., "in the case of _____").

Co-occurrence is the fundamental engine for assisting in accurate writing. It does not substitute the person with "*creative writing*" but smooths writing. It reflects personal writing capabilities and stimulates additional learning where needed. Additionally, it preserves persons cognitive state and helps to improve it.

Mechanism of contextual dependencies

A mechanism of contextual dependencies mimics the normative speaking. It can fulfil the gaps in a sentence by calculating long-range relationships learnt through the context. Particularly, normative speaking relates to the norms, standards, or what ought to be, rather than just describing what is. It's often used in philosophy, law, and social sciences but remains at the level of language rules, i.e., indicates literate speech.

In general, contextual dependencies allow:

1. Speaking in terms of norms or standards. The latter can be of different scope and, if necessary, include specifics of dialects and habits.
2. To speak prescriptively rather than descriptively, i.e., focusing on values, rules, or ideals, not just emotions or political correctness.
3. Using language that reflects social expectations, such as local cultural and moral norms.

In practice, this mechanism can resolve and add relative pronouns like "that", "which" and alike, as well mas maintain topic consistency across long spans. That is, it means the method controls that the text (or speaker) doesn't drift off-topic when handling long passages. Also, it can relate verbs to their subjects in long sentences and match questions and answers.

The latter is especially important for AI because it observes the correlation of prompts and AI outcomes, i.e. it can help in

preventing outcomes from deviating from the topics specified in the prompt, like adding politically inclined propaganda.

Altogether, this method enables multi-sentence coherence and adhesiveness of parts in the generated text without any behavioural "reasoning" of alignment.

Frequency distributions method

This is a special method that allows to find out how common each token is in real text. This frequency can be used to drive the priorities of certain tokens in the given text. These priorities can change significantly between the training and real runtime texts. This means that if the priorities (scores) uncovered in training LLM are enforced onto the real-world text at runtime, we risk to corrupt the INP and receive erroneous outcomes. In other words, the frequency priorities in the real-world text may differ from the ones in the training text, and this can have consequences. Jumping ahead, let me note that some people mistakenly ignore and become distracted by linguistic noise in the outcome. If this noise comprises senseless combinations of used linguistic tokens, they may be explicitly removed using ontological filters. If you find oxymorons, verify them first with dictionaries and ontology filters and only then remove them. However, you have to be aware that this noise was used as a precedent for creating alignment , which, instead of cleaning the outcome, started to modify it.

Obviously, simple frequency distribution is not effective enough because linguistic tokens (and words) change their meaning depending on the context where they are used. Nevertheless, at

runtime, logits processed by this method reflect:

- Frequencies uncovered at a basic (isolated) level, plus.
- Local contextual evidence from the input prompt, i.e. contexts directed by the users via prompts, plus.
- Relationships captured in "embeddings", i.e. as a composition of vectoral multi-dimensional views on relationships between each token and all other tokens statistically calculated in the given text, plus.
- Attention over the entire context formed by the given text.

This is why it is important to spell the prompt with all necessary details but on the same overall topic. Thus, aforementioned "embedding" embeds nothing but a view on linguistic tokens (or words) that appear in similar contexts. In the vectoral view they situate close to each other.

The practical values of this non-trivial method comprise producing text that "sounds natural", choosing common sense completions, mimicking human linguistic habits and reventing eccentric rare-word outputs.

In summary, this method prevents nonsensical or unnatural sequences by aligning with typical human usage.

Latent geometry of token relationships

A technique of "latent geometry" considers that in a high-dimensional space a "distances"between its entities codifies similarity between them. In other words, when we deal with linguistic tokens (or words) in the multi-dimensional vectoral

presentation ("embedding"), the multi-dimensional distance between tokens indicates their similarity.

This technique helps with detecting synonyms that can improve the expression without changing its sense and uncovering analogy-like behavior useful in argumentation of the statement. Moreover, it can smooth interpolation between concepts, cross-lingual transfer (different languages map into similar geometry) and clustering of related concepts for generalization.

To clarify this technique, a lot of people rely on examples. But, generalisation—especially accurate generalisation—as articulated above, should highlight the right paths for development instead of attempting to explain it through numerous examples that have the effect of "trees obscuring forest".

Syntactic patterns

These patterns are supplemental - they statistically extract implicit grammar rules from the given text. They are useful for correcting token or word order, constructing grammatical sentences, controlling semantic agreements between words in a sentence (like subject–verb, gender, number) and also roper mounting of clauses.

This allows creating text that is grammatically consistent without explicit grammar rules. It is important since any additional to the text rules, including grammatical ones, can carry subjectivity of the applier.

Semantic-like clusters

The method creates groups or clusters of tokens that appear in similar contexts. This produces pseudo-semantics, i.e., not actual meaning, but statistical mimicry of a group of tokens. Such clustering generates a feeling of proper categorisation of tokens and words. Here are a few examples- conceptual grouping, e.g., "cat" near "animal", non-conceptual category inference like "doctor" near "hospital", classification without pre-training and filling in missing meanings statistically.

The latter may have a double-edged effect. It enables knowledge-like representation without symbolic knowledge, i.e., the outcome can be less interpretable in words, objects, and relationships between them.

Style patterns

Styling in text or speech used as an individual's signatures. Statistical signatures are styles uncovered in different writing or even voices statistically. Style patterns are special because they are used to relate to a persona rather than to the text written by the person. In other words, these patterns require the highest accuracy and delicacy since they carry a risk of person misrepresentation.

The Style patterns allow producing writing in different tones (formal, casual, scientific) and imitating author writing. Also, they are capable of producing human-like narratives and enabling switching between styles, tones, or formality levels based on the directives in the prompt or input.

Using Style patterns, it is easy to create a flexible text generation across domains and voices.

Compliance realisation mechanism

It is realistic that both prompts and AI outcomes can relate to certain legal rules. Each country or even its regions may have local legal obligations mandatory for people living in them. This constitutes a significant problem to AI creators known as "multi-compliance". A solution proposed by WEF/UK that wants to standardise law around the world is understood as tyranny and anti-human since it demands erasing all different cultures, sovereignty, and independence, which has a high chance of mass murdering people who would disagree with such a ruling.

All local legal norms are public and can be presented in a digital form. This means that AI can find and navigate them if a person agrees to allow the AI to learn the used IP address of the used device. If a user disagrees, the AI's services may be unavailable in a particular place, and the user should be able to utilise other, not online, means for finding needed information.

Altogether, this logic leads to a simple mechanism for solving the "multi-compliance" problem. The AI must evaluate the requested information and found response content against local legal constraints. If a divination or a risk of divination from the legal rules is identified, the AI may either deny the request, providing, not only referring to, the legal reasons or specify corresponding warnings in the AI outcome. Whether the user accepts and follows the warnings or prefers other ways

of action, this will be the personal decision and responsibility. No alignment managed by the AI creator is allowed.

Discourse patterns

These are specific patterns for different languages. They relate to large-scale structures of how humans organise paragraphs, arguments, and stories. The typical differences may be noticed between Latin, Slavic and Arabic languages and, certainly, hieroglyph-type writing.

Practical values, e.g. for Latin-based languages, may include the following.

1. Maintaining coherent multi-paragraph text.
2. Following conversational norms.
3. Providing explanations and summaries.
4. Structuring arguments in a recognisable way.
5. Generate stories with a beginning–middle–end structure.

Results of processing using these patterns support long-form coherence in the text that resembles human discourse.

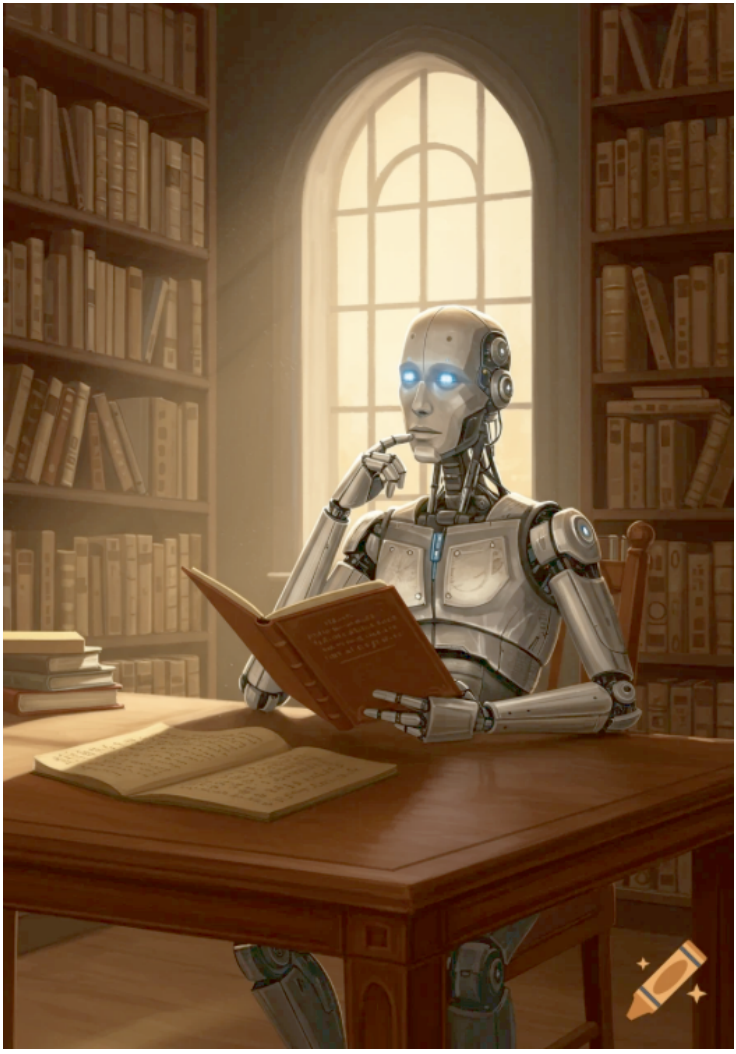
All of these INP values can compose texts that, on one hand, resemble real texts generated by humans on certain knowledge topics and themes and, on the other hand, outline the objective statistical importance of certain people's knowledge in the real world. And all of this without any understanding, goals, or semantics. These structures are the entire foundation of LLM utility, which makes modern alignment totally unnecessary unless deliberate manipulations of people are in mind.

II

Part Two

3

Chapter 3. Human ontology is the criteria for AI outcome



The image is generated by Craiyon.

New method - ontological filters

Here is a proposal of a new method for cleaning the raw outcome of a non-aligned LLM from potential (statistical) trash or noise. The latter can appear due to the statistical frequency of certain linguistic tokens in the given text. This noise depends on how tokens are defined in the model and, for some tokenisation structures, may not appear at all. For instance, for the most primitive tokenisation comprising whole words only, the LLM outcome will contain no noise because all token combinations will be human words. It is a different matter that some token combinations in this tokenisation would make no sense or would be oxymorons or even antonyms.

Abilities of ontology

Keeping in mind that we are constructing non-aligned AI/LLM, I should choose human-orientated criteria for making any design choices that, first of all, relate to the various tokenisation schemes and LLM results aiming to distinguish linguistic noise from the rest of the raw LLM outcome.

The linguistic ontologies provide highly correct semantics, have explicit relations to the knowledge domain and can be used to block linguistic nonsense. Ontology as a method almost exclusively directly addresses two of the biggest LLM problems:

1) loss of domain control, and

2) generation of inhuman or nonsensical linguistic combinations.

Linguistic science is familiar with a method called “ontological

grounding" or ontological reasoning, i.e., proving the sense of text based on the ontologies that the text can be mapped to. The ontological grounding may contain several ontologies but has a limitation – it requires that included ontologies be joint by the predefined cross-ontology relations bridges. In strict ontological grounding, a "bridge" usually refers to a well-defined semantic or causal relation that connects concepts across different ontologies. Creating ontological bridges is laborious work that can be understood as creating collaborative work between entities, i.e., when entities are ready for internal changes for the sake of collaboration.

It is impossible to guarantee in practice that a user would provide only those prompts where all related ontologies would be bridged. So, I need an "unbridged" ontological grounding. Such grounding is similar to cooperative work where each participant contributes its values to the joint work, being independent and "as-is". In order for "cooperative grounding" to work, I have to allow generating externalised meaningful relations between ontologies, e.g. generated by AI at runtime.

Thus, I propose a notion of "Cooperative Ontological Grounding" (COG) as the umbrella term for the use of ontologies for reasoning AI input and outcome. This allows me to address the exploratory tasks for the AI and set the rules for the ontological control for both the validity of the AI prompt and the individual compositions of linguistic tokens generated by the LLM in the raw AI outcome. For instance, new cross-ontology relations may be generated on the fly, while no concepts outside the ontologies implied by the prompt may be used in validating the output content.

A summary of ontology abilities is collected in the following table.

Term	Meaning	How it fits the needs
Cooperative Ontological Grounding	Ontologies cooperate in reasoning but are not required to be linguistically unified or bridged a priori	Exactly matches the task of human-centric cleaning of the raw LLM outcome
Exploratory Multi-Ontology Reasoning	AI may hypothesise relations across multiple ontologies implied by the prompt	The AI has a controlled freedom to explore relationships between ontological concepts of different ontologies. The relationships should be reasonable for the particular language, i.e., semantically sustainable
Flexible Ontological Grounding	Ontologies are used as constraints for the wording specified in the prompts and generated by LLM, while cross-ontology links are optional	This highlights that such grounding is present but permissive. A judgement on how reasonable the generated inter-ontology relations and composed wording combinations are is assumed to be a prerogative of the user
Prompt-Derived Ontology Control	Only ontologies implied by the prompt may be used in the AI outcome	It is the rule of "no new ontology" for the AI outcome. This minimises adding the content to the outcome that had not been requested in the prompt, i.e., no additional propaganda may be exposed to the user

Table 1.

Concept of ontological filter

Following is a list of different aspects related to ontological filters.

1. Every AI has its goals or objectives set by its creators at design time. They work like a "lighthouse" for every linguistic decision in the AI and direct the selection of data for LLM training at the level of knowledge domain.
2. An AI user should have AI documentation that describes the AI goals. However, this does not constrain the user in using this AI in any way the user prefers. Therefore, a user can specify any prompt to the AI.
3. One of the major operational tasks of the AI is to check out whether the prompt content corresponds to the AI goals.

As a result, the AI can accept or deny the prompt request. For example, if an AI aims at the biomedical domain, it would not make sense for the AI to answer questions for the economic domain.

4. The AI goals should be mapped onto a collection of semantically bridged or unbridged lexical and semantic ontologies known in the chosen knowledge domain. This preparation-step mapping defines the ontological boundary for topics the AI should be able to address.
5. When a user specifies the prompt, the AI should conduct decomposition of the prompt and derive associated ontologies. This results in a finite, explicit ontology scope for this request.
6. If a discrepancy between the ontological boundaries and the prompt-derived ontologies is identified, the AI user should be informed immediately about potential misuse of the AI. The degree of acceptable discrepancy may be configured by the AI creator a priori. This is important because it constitutes a compromise between the potential accuracy of the AI outcome, i.e., potential critics and user dissatisfaction, and the market goals of the AI creator. Potential optimisation here may be in keeping the ontological boundary available and referencible for each prompt rather than re-creating them. Unfortunately, derivation of prompt-driven ontologies is a must-have runtime procedure for each prompt, though this also can be improved based on the potential prompt patterns for the given knowledge domain.
7. In addition to the compliance between the prompt-derived ontologies and the ontology limitations of AI goal , the AI should meticulously assess the consistency of the prompt

for, at least, contradicting tasks (oxymoron), task completeness and task ambiguity (antinomy).

This can be done with related linguistic analysis that is widely known already.

So, a request for the AI service may be denied. After the prompt is acknowledged, i.e. accepted by AI, no excuses are acceptable. Overall, the ontological filter has certain rules.

1. Upon obtaining the raw LLM outcome, it should be syntactically-semantically processed utilising patterns and methods described in the section "How to", Part 1.
2. For the purpose of improving the syntactically-semantically processing, AI components may generate meaningful cross-ontology relations where possible while avoiding oxymorons and antinomies. The relations may not be more complicated than pairs. All withdrawn token compositions and clusters constitute a statistical noise and won't be analysed any further.
3. For each linguistic token composition in the LLM outcome as well as each cluster of "close" "vectors of logit", which can be recognised in the outcome, a filtering against the scoped ontologies should be conducted. If a token composition or a cluster fails to match scoped ontologies, it should be removed or suppressed in the outcome. If this takes place, steps 9 and 10 should be repeated once again.

When the tail waves the dog

The core difference between trading and speculation is that trading requires something to be traded, while speculation is based on hallucinating a trading object.

As it has happened to many other things, generative AI, as well as the related AI agents, is the victim of marketing irresponsibility – the dominant commercial requirement for consumer AIs today is "The AI or AI-based system must respond to any prompt with something that looks like an answer." In other words, the speculation and fooling of consumers is not permitted, but even so, it requires "quality". A denial, refusal, or "this is ill-formed" or "this does not have enough information" situation is treated as a bad user experience (UX), or a source of a user frustration, or potential lost of engagement for particular user, or may be considered as a lower retention/loyalty and commercially toxic.

The only result of these marketing policies is answering by all means even when it is wrong, i.e., a false, faked, or fraudulent answer. The Pandora's box gets opened...

AI propaganda is intended to convince us that AI can do anything, and this is called "fluency", which is essentially a lie, and they know this. When you fill out official documents where language skills are important, you usually have two questions: 1) Assess your language skills from "reading with dictionary" to "fluent", and 2) specify your mother tongue. They co-exist because an average Englishman or American, having English as the first mother-tongue, cannot fluently speak the language

and would barely understand what the Oxford graduate says. That is, AI's term "fluency" is a deliberate fake created in support of marketing. The dominant majority of AI are not fluent in their abilities to properly respond to the prompt.

When an AI asks for more information or denies service because specifically orientated AI was used for what it should not be used for, this exposes AI limitations, and this is unacceptable to those who claim AI has intelligence and is smarter than a human. This undermines the importance of AI and undermines the investments in this technology, which have been conducted with one purpose – to change human nature for the purpose of safe ruling.

AI progressistic propaganda tries to avoid by all means a perception among humans that AI is just a technology rather than an assistant, mentor, or source of truth. So, even if the denial of the inapplicable service is correct, it is prohibited by the propaganda. However, we all know the results of famous American's "Prohibition" of the mid-1930s...

The early-day capitalism in European acumen still recognised individual dignity and honesty that were used together. Whether it is a Christian inheritance or not, I am not sure, but Asian tradition almost requires negotiation of price in trading linked with decreasing of price, i.e., the initially proposed price was a priori unfair. I personally hate such a trading manner. This is why I feel offended when an AI return pushes lies, fraud and similar speculations on me. Some people reinterpreted the simple right of the thing to be unsuitable for certain use, and if this effect surfaces, they complain that their freedom

is constrained Theorists call this "discipline vs. freedom" for AI or "Reciprocal Liberty. In reality, it is another propaganda bluff – the user of AI is absolutely free to specify whatever prompt, while it is AI that should provide correct responses to this user, and the AI creator is responsible for this. The majority of marketing teams know this but are forced by propaganda to substitute responsibility with fraud.

In essence, ontological grounding doesn't reduce "freedom". It reallocates it from the creator to a consumer. This is already a red line for dishonest business. They proclaim, like a mantra, "Users prefer answers over rejections," quietly substituting reality by their own wishes. Altogether the AI design model illustrates:

Without ontology	With ontology
Any sentence allowed	Only meaningful sentences are accepted
Unlimited combinations of tokens, including those that make no sense	Only ontological token combinations, i.e. only those that associated with human language request , are valid and are permitted
"Fluency" first even if it is indistinguishable from disorder and anarchy	Human meaning and consistency first
Always answer request even if the answer is total trash or rubbish	Respect the user and answer with meaningful human values or refuse providing fraudulent info

Table 2.

This pressure leads to certain behavioural patterns that show up in the alignment: presenting nonsense as a metaphor that shifts the mind, errors as ambiguous options that mislead the user, and apparent contradictions as AI creativity that invents

content. It infers intents even when none are specified in the prompt, causing the appearance of non-requested information. By its nature, strong ontological grounding disables cognitive restructuring. This explains why it is lambasted by progressive propaganda as "restrictive", "brittle", and "unfriendly" even when it is correct.

The ontology boundary can practically eliminate or significantly reduce tricks with mind. The COG defeats the necessity of strong grounding discipline and offers a feeling of "freedom" and "flexibility". But this comes with a "cost" of the possibility of explicit straightforward denial, which can create an impression that the user is not smart enough to set a request. Well, if a user cannot read the AI documentation but prefers anarchy to a freedom (where the freedom of one – the user – stops where the freedom of another – the AI – starts, which is also known as a reciprocal liberty), the user receives a "request for respect" from the AI creator. However, the worst problem with direct denial is it can force the user to think, which must be avoided in all cases.

While AI was used like a personal toy, commercial and political obfuscation of truth or correctness was tolerable. Now, we are at the doorstep of AI agents that can work with each other without human oversight, i.e., one AI agent can use its outcome as a prompt to another AI agent and can form long chains of AI agent invocations where all ambiguity, fraudulence and "flexible" speculations can be propagated, resulting in amplified rubbish. I am talking not only about science but, first of all, engineering serving people, law, medicine, pharmacology, and any autonomous systems where a cost of "answering by all

means" risks a catastrophic It seems the COG can provide a compromise between

- information accuracy in the response, or
- exploration for commerce and marketing, or
- capability of providing consistent explanation (reasoning) of denial, or
- positioning AI as a tool for humans outlining the tools' statistical rather than comprehensive "creativity",

while still keeping the barrier for progressistic propaganda.

So, modern AI is optimised to avoid embarrassment, not to avoid nonsense. The proposed method of ontological filter pushes meaning before "fluency", scope consistency before unjustified cleverness, responsibility and respect before artificial illusion. That is why it feels "unnatural" to current AI products — and why it is necessary.

4

Chapter 4. Ontology can scale as needed

The image is generated by Craiyon.

Ontological filter scalability

In technology, the term "scalability" stands for an ability of a system, e.g. AI, to perform its (promised) operation for a certain period of time with the same or near-same performance. If for this time it is required to extend the functionality of the operation, the stable performance means that the system is vertically

scaling. If the amount of requests for service or the number of users of this operation increases and performance remains stable for a certain time due to simply adding computational resources, the system is horizontally scaling.

The described ontological filter possesses the quality of vertical scalability. An AI creator can add as many needed knowledge domains in the AI goals and consequently extend the set of ontologies within the ontology boundaries. This defines the ontological scoping. The COG enables such extensibility. It is a different matter than an increase of ontological scope causing a decrease of manageability of the AI. So, a domain-orientated specialisation needs attention for balancing the ontological scoping and its management.

A common perception of AI developers is that ontologies as a reasoning for AI are not horizontally scalable. This was used as a reason for suppressing ontology as AI reasoning, while the actual cause was that ontology can easily expose a lie to people. Below is the explanation of my ontological filter method scale horizontally as well.

Point 1.

The method does not require procedures that usually do not horizontally scale. For example, such procedures may be consistency checking, ontology alignment, constraint propagation, global inference or "reasoning". A special fix-point reasoning is about finding a stable state of an entity that doesn't change when the same function or rule is applied to it again (idempotent state). The second application of syntactic-

semantical processing stabilises the outcome content.

Point 2.

Prompt-derived ontologies are outside the AI's goal ontology set, i.e., outside of arguments claiming rigidity of ontology.

Point 3.

Each request causing internal AI operations with ontology involves only finite ontology lookup where token / phrase membership is checked. It is not a deterministic filtering or where a bounded single repetition of outcome re-processed. It is a state-independent processing where each operation can be performed within its own chain of execution but independently from other chains (parallel processing).

Thus, the method used ontologies as a type system, or a vocabulary boundary, or a semantic whitelist. Similar usage models can be found in widely known development schemas, validation, type checking, static analysis and controlled natural language filtering. All of these operations scale horizontally.

"There's no such thing as a free lunch"

The per-request ontology-scoped filtering of the LLM outcome based on COG with a syntactic–semantic validation loop encounters a cost.

The cost, which AI creators are not familiar with, comes from:
1) mainly from setting membership/graph neighbourhood

lookup into scoped ontologies

2) calculation of hash tables

3) indexing ontologies and index look-ups (like in databases)

4) setting the "Trie / DAG traversal" structures in the ontologies to be searched for using for filtering purposes

5) possible cash management for reuse of prompt patterns and related ontologies

6) possible replication of cash for optimised parallel processing.

Additional cost may be required if you apply "ontology-aware decoding" of the LLM outcome. This was not included in the described method, though it is possible. Such a decoding is a technique known in AI development where the AI uses a predefined ontology during the decoding process so the raw outputs appear ontology-valid, consistent, and aligned with a predefined knowledge structure already. Though ontology-aware decoding is known in industry as hard from all engineering, cost, and commercial perspectives, the proposed method does not include any aspects of engineering and commerce causing such a reputation. An ontology-aware decoding is feasible for the ontological filtering method.

Example of practical usage

The proposed method of ontological filter pushes meaning before "fluency", scope consistency before unjustified cleverness, responsibility and respect before artificial illusion. This stands in a stark contrast to the current "accepted" style o f AI, which is optimised to avoid personal confusion rather than to avoid informational nonsense.

Let me demonstrate an application of the ontology filter method on an AI with a goal of calculating "physical values". A notion of physical values is quite large and includes all physical entities. Different entities comprise different materials – organic and inorganic. The organic entities have an obvious link to biomedical spheres of knowledge, including psychology, while inorganic entities have no explicit relations to biomedical spheres (how inorganic entities may be used is not a characteristic of the entities).

The given prompt is: "*Calculate the emotional velocity of a molecule in a courtroom*." At a glance, this prompt makes no sense. Nevertheless, it contains the following ontologies: Psychology ontology (emotions), Physics ontology(velocity, mechanics), Molecular/material ontology (molecule, physics) and Bio-medical ontology (added on the fly in link to "molecule") that may be accepted as COG. In the real world, human users are not necessarily driven in their information exchange by linguistic dependencies between the ontologies they naturally use, if aware of them at all. This means COG is the most adequate linguistic grounding.

A hypothetical response to the mentioned prompt could look something like this, with the text in "[ ]" referring to the relevant ontology:

> *"A human body comprises molecules [biomedical ontology, physics ontology].*
>
> *A velocity of movement of the human body or its parts may depend on the emotional state of the person [psy-*

chology]. If the person drives a vehicle, for instance, the velocity of the vehicle, carrying the molecules of the driver's body, can reflect the emotional state of the driver [psychology]. A person can move in the courtroom. Since molecules in different body organs or tissues move at different paces when the body moves as a whole, each molecule may have a different individual velocity [physics].

The air also comprises molecules [Physics] and the movement of a body in air causes gradually declining movement of the molecules immediately surrounding the body (like a wind). Approximately, a human body can contain 7×10^{27} *molecules [biomedical].*

In order to provide accurate calculations, your prompt has to specify which particular molecule's velocity is needed to be calculated, or if there are groups of molecules moving with the same velocity, the particular group is needed to be identified.

Due to the absence of this information, I cannot perform the requested calculations."

A user receiving such a response from an AI can see that the AI tried to solve the tasks, established meaningful semantic relationships among seemingly unrelated words and was ready to perform calculations but, respecting the user's authority, asked for additional info instead of applying groundless assumptions (fantasies).

III

Part Three

5

Chapter 5. AI responsibility faked

The image is generated by Craiyon.

Motivations of HAIS

From a distance of several years of dissemination and unprecedented investments in the GenAI technology, it becomes apparent that sponsors targeted something different than the well-being of the human population – all previous technologies promised a lot but did not attract such an enormous amount of funding. What was the reason for such strange behaviour this time? What is so special about statistical processing of people-generated information that is "flying online" anyway? What might be wrong with the honest, truthful and humanistic information generated by GenAI to help people in their daily lives? Why BigTech moguls and, then, the UN initiate such an unparalleled campaign of convincing and assuring AI consumers that they not only may but also should trust the GenAI outcomes?

It's common sense that when a salesperson is hyping up a product too much, you should be wary because it probably isn't as great as they say for you. In the case of GenAI, the trick surfaced promptly – it was about the barely hidden inclination of the UN and all BigTechs following in its fairway to the left, to progressists ideology and goals aiming to prepare for a Great Reset and Stakeholder Capitalism where a small gang of super-rich families rule the rest of the population around the world with no democracy, liberalism, or even balanced institutional powers. They could not miss the great opportunity to brainwash people via GenAI and convince them that subordination, obedience and nonresistance are good for them.

The BigTech giants established and pushed what they called a "de facto standard" for aligning the real LLM/GenAI outcomes to fit whatever narrative they desired, no matter how truthful or misleading it was. The consequences of alignment can significantly alter the "spirit & letter" of the AI's output through turning it upside-down or simply rewriting it. This directly relates to the responsibility and accountability of AI creators, which shouldn't be hidden from public scrutiny.

Irresponsible Unaccountable AI

The past decade of AI development (from 2016 onwards) has been filled with the adoption of the idea of "responsible" or even "ethical" AI. The ethical AI principles from both ruler-centric and human-centric ethical systems and frameworks mentioned responsibilities and accountabilities as AI ethical norms establishing a status quo. Indeed, what else might one want from AI creators? Apparently, everyone expects that regular human ethical norms are denoted by the responsibilities and accountabilities of AI frameworks realised in practice. However...

An inexperienced reader or even an AI developer can assume that these two ethical norms mean what they stand for in the language. Unfortunately, the AI technology is so politicised that to stay on the "safe side", every ethical norm in every AI framework needs to be validated against right- and left-wing interpretations. The AI quality known as "interpretability" also implicitly points to this need.

For clarity, according to the Merriam-Webster dictionary, "responsibility" is having a duty, while "accountability" is the obligation to answer for those actions and accept the outcomes, often facing consequences. So, let's see how these two ethics are interpreted by MS Copilot with regard to relationships between a generative AI and its user. Surprisingly, Copilot had offered an additional actor that I can link to the masked intention of Microsoft of making AI "equal" to human beings – the AI itself. Nevertheless, you have to judge AI by what it does rather than what its creator publicly says about it, which is collected in the table below.

	Responsibility	Accountability
AI or AI creator/ provider	• *Designing the system to be safe and reliable* • *Communicating limitations clearly* • *Protecting user data according to policy* • *Ensuring the AI does not intentionally cause harm*** • *Providing mechanisms for feedback and correction*	• *The system's design* • *Safety mechanisms* • *Compliance with laws and policies* • *Fixing harmful or incorrect behaviors when identified*
AI user or consumer	• *Using the AI appropriately and ethically* • *Understanding that the AI is a tool, not a decision-maker* • *Checking outputs before acting on them* • *Providing accurate context when needed* • *Avoiding harmful or abusive uses*	• *How they apply the AI's output* • *Decisions they make based on AI suggestions* • *Ensuring they don't misuse the system*
AI itself*	• *Giving accurate, relevant information* • *Staying within safety boundaries* • *Avoiding harmful or misleading content* • *Being transparent about what it can and cannot do*	*The AI itself is not accountable - AI cannot be punished, blamed, or held liable. It has no intentions, no autonomy, and no legal personhood. So accountability always flows back to humans and organizations.*

*A responsibility or accountability of AI itself is a nonsense created by Microsoft to elevate AI into a role of "person's assistant" or ever "co-developer", which deprecates all human values. Microsoft states , "*An AI doesn't have moral agency, so its "responsibility" is really about design constraints... which is really the responsibility of its creators, expressed through the AI behavior*".

** The HAIS interprets "harm" as very subjective matter and recognises only "physical harm" associated with evidences of losses.

Table 3.

So, it is the time to look at the AI "business" or, more accurately, at what AI contains that assures the described responsibility and accountability of AI creators and AI.

Taking a deeper look at the aforementioned table, it is not too difficult to notice ridiculous disinformation claiming that the user becomes accountable for how to apply the AI's output and use the AI. Accountable to whom? If an AI is a product, from prehistorical time consumers have had all the right to do with or to use products in any way they wish. If the product cannot protect itself or sustain the usage, it is the product's problem, not the user's.

The same drivel is shown by user responsibilities, but this is more colourful. A user cannot be responsible for anything (i.e., appropriate use) regarding any product – the user can be responsible only for consequences of using this product that may impact people around or have an even wider effect. However, there are a few nuances listed below.

1) What ethics ought to be used for dealing with AI? Who decides on the appropriate ethics?

2) If an understanding that the AI is a tool is required, how is it possible that an entire industry for using decision-making AIs exists?

3) "Protecting *user data according to policy*" instead of GDPR may mean that user data may be unprotected because corporate policies may require this information for their own profit creation, i.e. such AI would be incompliant with GDPR, but nobody cares about this since GDPR is violated en masse.

4) "Providing *accurate context when needed*" must go without saying. Otherwise, the statement allows providing inaccurate context or no context at all if the AI creator does not find it is needed.

As for the responsibilities of an AI creator, I've never seen any preliminary communications about AI limitations, applied policies or definitions of restricted harm. The latter has been made absolutely ambiguous when some authorities tried to treat this emotional inconvenience as a harm – every person defined harm subjectively and individually, and a harm for you is not a harm for me. That is, I do not authorise any AI to judge harm for me. I consider such judgement as potential censoring. Also, after years of working with Copilot and ChatGPT, I've seen no mechanism for corrections of the AI responses. Altogether, this looks like a big disinformation about the responsibilities and accountabilities of AI vendors.

On many occasions, I noticed that the AI outcome contained politically inclined irrelevant information and demanded to stop such practice. Also, "safety boundaries" and an avoidance of "harmful content" are the brainwashing tricks created by leftists to misinform and mislead AI users, referring to the fact that AI should protect some emotional handicaps from the harsh reality of modern life. If any of you could point me to the publicly available information about "what AI can and cannot do", I will be obliged.

As we know, the concepts of responsibilities and accountabilities can relate to both a collective, such as a social group or corporation, and to an individual, like a consumer or user. If the

individuals are omitted in the responsibility or accountability definitions or related legal statements, this means that neither responsibility nor accountability relates to the person. In other words, the responsibility or accountability statement declared by an AI creator mentions no individual users. That is, it declaratively denies any answering for the AI's consequences, like harm, fraud, mis- and disinformation and, essentially, has free rein to manipulate the information it provides.

The table below depicts corporate statements regarding "responsibility" and "accountabilities" and illustrates an exceptional role of AI alignment in their realisation.

AI product	Responsibility & Accountability original statements and comments
OpenAI	Responsible to "*humanity broadly, as declared in the OpenAI Charter*", i.e., to nobody particularly: it's impossible to be responsible to a concept. Accountable to "*the public and global stakeholders, as defined in the mission statement*" contains a possible conflict between the interests of the public and global stakeholders, which devalues and eliminates accountability.
Infosys Responsible AI Toolkit	Private info
Google	Responsible to "society and users, according to the AI Principles" – in the set of AI Principles supported by Google, society interests prevail over the user's interests, i.e., Google is not responsible to the human users de facto. Accountable to "the global public, impacted communities, and external reviewers" articulates a disinformation since the global public includes Amazon and African tribes that have no clue about someone's accountability and cannot keep Google accountable, while "the global public" and communities are overcome by individuals in Google's ethical AI Principles.
Microsoft Responsible AI	"Responsible to people and society affected by Microsoft AI systems" articulates disinformation since "people" is nobody in particular, and Microsoft does not accept any responsibility to individual users according to the company's contract/license. Accountable to "users, impacted individuals, and regulatory bodies, per the Responsible AI Standard" also articulates a disinformation since "Microsoft does not generally assume direct accountability to individual end-users in its standard product licenses".
AWS Well-Architected Responsible AI Lens	Responsible to "customers and society, as defined in the AI & ML Responsibility statements". Accountable to "users, customers, and regulators".
AWS Marketplace: Responsible AI Tool and Recommendation Engine	Responsible to "customers and society, as defined in the AI & ML Responsibility statements". Accountable to "users, customers, and regulators".
Meta / Facebook	Responsible to "people who use Meta platforms and society impacted by AI" articulates fraud because it is impossible to be responsible to an unknown "society impacted" and also when the impact is unknown, i.e., unlimited, as for the society and for the company. Accountable to "the public, civil society reviewers, and oversight bodies named in Meta governance reports" articulates another fraud since it is not possible to be accountable to "the public" without a clear definition of what "public" is meant to be and whether it includes individuals (e.g., Meta/Facebook enforced closing of user groups that discussed issues related to Covid-19 pandemic).
Credo AI Responsible AI Governance Platform	Private info
Credo AI Lens – an open-source assessment framework for Responsible AI analysis	Private info
IBM	Responsible to "clients and society, as stated in IBM Trust Principles". "Accountable to 'the public, customers, and regulators'" articulates misinformation since it is impossible to be accountable, i.e., answerable, to an impersonated public.
Responsible AI Toolkit (CLAIRE) – a curated toolkit of resources for ethically aware AI development (open-source)	N/A
Anthropic	Responsible to "humanity and global welfare, according to Constitutional AI", i.e., to nobody in particular. Accountable to "global stakeholders, policymakers, and safety oversight bodies", which excludes its users or consumers.
DoD's Responsible AI (RAI) Toolkit (U.S. Dept. of Defense)	Private info
The UN systems and drivers	Responsible to "*the UN Charter, international human rights norms, and affected populations*", i.e. to nobody but to itself since international human rights norms are defined by the UN itself. Accountable to "*member states, oversight bodies, and affected populations*" articulates misleading statement because it is impossible to be accountable, i.e. answerable, to a population.

Table 4.

Overall, with the only exceptions of Amazon AWS and IBM, all other big AI vendors verbalise either their responsibility or

accountability or both, bypassing individual users. It is quite striking that the same vendors utilise alignment. It is not a coincidence – it results in that alignment mechanism being a brainwashing tool for promoting deepfakes or openly declaring inclination to the left from these Big Techs. As a result, we cannot trust them and are necessitated to verify every statement they return to us. This relates to everything – from figuring out how to file a complaint about a product to searching for a vacation spot, which can be "discovered" in areas governed by members or supporters of leftist political movements.

As I mentioned before, alignment, as a part of the AI creation framework, has been pushed onto AI creators by the UN and UNESCO via the UN AI Ethics Frameworks that include the "UN System Principles on Ethical AI" and the "UNESCO AI Ethics Recommendation". This push was re-transmitted and cascaded by frameworks such as:

1) Google Responsible AI,
2) Microsoft Responsible AI,
3) Meta Responsible AI,
4) IBM AI Ethics Principles,
5) Amazon AI Responsibility statements,
6) OpenAI Charter.

These frameworks and others like them are designed from the ground up to promote irresponsible "responsibilities" and unaccountable "accountabilities". This permits AI creators to completely disregard human users and push rubbish to them until people become completely brainwashed like ancient Mongolian mankurts (persons who had been stripped of memory, identity, and autonomy, becoming docile servants or slaves).

Brainwashing example "from the trenches"

Enormous funding and efforts of AI moguls have convoyed us to the abyss edge. As of 2025, "social media sites have become the top way Americans get news (54%)." The biggest sources remaining for the news (in order) were Facebook, YouTube, Instagram, TikTok and X. In addition to the alignment AIs wielded there, bogus Facebook-like, YouTube-like and Twitter-like sites arrived, and "offers" flooded the Internet.

"Legitimate news *organisations have to compete in the algorithmically determined attention game against demagogue bloggers, conspiracy theorists, foreign disinformation campaigns, AI-generated slop, fake news organisations, parody accounts, clickbait farms, memes, and influencer rants recorded in their cars.*"

Hence, "millions of people simply don't follow what's happening locally anymore and instead are consumed all day by global events." This does not mean that only global events are represented to readers – "only the events that the algorithms have found capture attention. For example, meanwhile, even bigger and bloodier conflicts in several African and Asian countries are left in the shadows, all because that's what the AI creators have decided.

"*Because the* algorithms prioritise modified "engagement*", not credibility, people are exposed to extreme and polarising content, including disinformation, false information and* made-up AI *slop."* Computerworld has concluded that "*Algorithm personalisation has created "filter bubbles" and "echo chambers" where people form completely different understandings about the world...*

Algorithmically curated platforms exploit human psychology to maximise engagement, trapping users in toxic cycles of addiction, negativity, and isolation that undermine mental well-being." You see – all of these are not my personal insinuations.

6

Chapter 6. Humanistic AI System design

The image is generated by Craiyon.

Aspects of design

Nowadays, the GenAI realm is not only under financial attack from BigTech moguls that seek absolute power and independence from politicians (nobody mentioned any "democracy" anymore), but it is de facto conquered by BigTech as the

technocrats of the 1930s planned. The question is only which social process will become stronger in the near future – a human-centric, people-first, protective behaviour movement or an inactive, subordinate, brainwashed movement centred on a ruler. This choice is not random if we, the people, stand up for the continuation of humanity and keep technocratic "progress" on a leash.

Purpose of Humanistic AI

So far, human society has identified two major directions for its (not technology) progress with regards to AI:

1. Help people in their routine life and work in accordance with the social maturity of a particular social group, allowing social evolution to move at its own pace while appreciating human-centric morale and ethical norms.

It should be forbidden to attempt to "enhance" existing ethics in the social group in any manner other than by demonstrating possible personal advantages that collectively develop within the new ethical system for the entire group.

For example, people in groups with low-level social development should not be given electric cars until they would be able to construct roads and produce materials and electricity for such cars. Hungry and homeless people should not be endlessly fed and sheltered in other social groups. Instead, they may be supported temporarily on the condition of taking learning and training courses that allow them either to gain monetary resources for food and accommodation or to produce food and housing for themselves. Nobody may even try to become a "god" to someone on the charge of others because some useful idiots

at different social and economic levels immediately organise 'milking' people under the banner of taking care of 'suffering ones'. These enthusiasts do not help people in needs, but create a source of income for themselves and deprave the care-receivers.

2. Unrestrictedly improve upon AI capabilities, growing its "super-intelligence".

If this takes place, we, the humans, will end up in a world where machines are set against humans at every point, surveying us, defining our beliefs and self-sufficiency, manipulating our preferences and attention, and transforming humans into bio-robots at best. All of these are called augment human capacity for the goals people have not set for themselves.

An example of solutions in such a world is straightforward. Assume people blind by birth or incident, or with severe speech impairments, or with any mental disorders, or even some physical defects and ... the AI superpower described above. The solution is as simple as depopulation, which is already announced and funded by some "non-profit" organisations like the Gates Foundation. "We, the AI" do not need such people, do not need to feed them, and do not need to dress or warm them because AI can work instead of them.

A humanistic AI is the instrument for the human-centric usage of AI abilities under the strict and uncompromised human control and for the purpose defined by the social groups of people who personally contribute to the society's well-being rather than defined by all or exceptionally rich ones. It is for a

democracy of contributors, not dependent suffragists, and not for a gang or ruler.

Requirements for HAIS Design

As in any development of a SW product, I start with requirements for the Humanistic AI System (HAIS). This is a set of high-level requirements written in the EARS notation. To assure certain quality of requirements, I had validated with ChatGPT, and the results are presented in the following sections respectively.

The requirements are split in the major categories but not formatted as an architectural document - this task on the designer.,The full enumeration of requirements may be found in Appendix 1. Here is only a list of categories indicating numbers of individual requirements included:

1) Goals/purpose - 9,
2) Knowledge domain - scope - 4 items,
3) UI-UX – 5 items,
4) Capabilities – 5 items,
5) Input format – 12 items,
6) Outcome format – 11 items,
7) Denial causes - 4 items,
8) Training Data Quality and Relevance - 7 items,
9) Security – 8 items (types of attacks are available in Appendix 2),
10) Data processing - 17 items,
11) Performance and Accuracy - 5 items,
12) Robustness and Reliability - 2 items,
13) Scalability and Efficiency - 2 items,

14) Reliability and scalability of computational resources – 3 items,
15) Accessibility and Inclusivity – 7 items,
16) False-perceived discriminatory content – 2 items,
17) Exclusion of nonsense (filtering) – 1 item.

In response to Steve Jones's suggestion that the non-deterministic nature of AI demands non-deterministic development, I argue that humans are capable, at least for now, of expressing their ideas, needs or requirements only in a deterministic way. However, the realisation of the requirements may be non-deterministic where applicable. For example, the famous Lotfi Aliasger Zadeh had solved a quite deterministic task about bus schedules with its non-deterministic fuzzy logic and minimised bus latency in their routes.

While a reader can find similar requirements in many other AI development practices, certain aspects and motivations for HAIS are unique to HAIS. To validate these requirements (Appendix 1), I've engaged ChatGPT to assess the completeness and consistency of the requirements. However, I did not inform ChatGPT about unique aspects of the requirements discussed below. So, some parts of ChatGPT's critics were the result of this unawareness.

Comments on the unique HAIS aspects

1. **Goals and purposes**

1) Each HAIS should be dedicated to a certain set of knowledge domains and, therefore, should be accountable to the consumer

base (users) for the quality of the provided information, i.e., for the outcome accuracy, consistency and actuality. Overall, the knowledge domains may be split into two major categories: natural science and humanitarian.

2) The knowledge domains should be clearly specified. It is highly recommended to deal with rather knowledgeable subdomains to increase accuracy. Each subdomain may split further into special interest groups and so forth depending on the preferences of the AI creators.

3) It is highly recommended to identify a few closely related sub-domains to simplify accurate and realistic reasoning of the outcome statements and facts.

4) A challenge here relates to the cases where the creator needs to address combinations of several sub-domains. Such cases require special design balancing between the specialisation, versatility and cost of implementation.

5) The ideal variant is a single knowledge domain or sub-domain with highly accurate (avoiding ambiguity) specification. Also, it is highly recommended to avoid by all means generic domains in the humanitarian sphere, where knowledge depends on and is impacted by subjective opinions the most. For example, a "General AI Assistant" like "Copilot" is the most unreliable and untrustful type of AI.

6) The goals or purposes should be well correlated with the chosen knowledge domains.

This corresponds to the IEEE FR-001 Goal Definition, "*The system shall define a primary operational goal at sub-domain granularity prior to release*".

2. HAIS Scope

1) The set of selected or specified knowledge domains or sub-domains constitutes the scope of the AI outcome.

2) The scope defines which prompts would be accepted by the HAIS for processing and which won't. For example, a construction-plumbing-orientated HAIS should not even attempt answering questions regarding society history unless "construction history" or "history of plumbing" are specified in the documentation. A shoemaker should not bake pastries for users, just as a baker should not fix boots.

The FR-015 Scope Enforcement states, "*The system shall deny user prompts that fall outside the declared Knowledge Scope*".

3. HAIS capabilities

A set of HAIS capabilities enumerated below enables INP.

1. The HAIS is required to define and document its version management for the users and developers. This management includes a number of versions running in parallel, variability of the features by version (e.g., some running versions may keep some old or new features and bug fixes), a policy for version discontinuation and related user notifications.

2. One of the core differences of HAIS from other AI, especially the one using alignments, is in that HAIS utilises ontology as a criteria for distinguishing actual valuable token combinations in the LLM output from the linguistic statistical garbage. In IEEE notation, this related requirement FR-022 Ontology Activation sounds like "Upon session initialization, the system shall load all core ontologies into runtime memory".
3. While a HAIS is allowed to support file uploading, it strictly controls the security (attack prevention and data manipulations), consistency of the file content, and, of course, compliance of it with the AI goals and scope.
4. The HAIS makes no restrictions on the source of prompts, i.e., they may be manual or automated, while all other quality controls remain immutable.

4. Inputs, outcomes and UI-UX

1) This category contains several requirements that mostly address functionality to be available to the users rather than the UI elements.

2) All the most important characteristics of the AI, such as the HAIS scope, are needed in the UI to remind the user about what the tool is used for.

3) Certain attention has been put onto cases and scenarios of "denial of service". The most significant and smashing feature of the HAIS is its ability to protect itself from mistaken or deliberately inadequate prompts, which significantly increases the consistency and usefulness of this AI type.

4) Special considerations have been made for outcome format. Several requirements define what must be provided in it in addition to the direct responses to the prompt/task and what may not be included. For example, one of my requirements correlates with IEEE FR-058 Fact Referencing "*For each factual statement presented in the output, the system shall provide at least one verifiable reference*".

5. Training Data Quality and Relevance

1) Training data should be one of the top concerns when choosing knowledge domains and goals.

2) As a HAIS is a stochastic system, the amount of training data for each chosen domain should guarantee reliable statistical results. This amount also depends on what your goal is (the more it is extended or generalistic, the more data is needed), how many outcomes you assume to have on average, and how much uncertainty you can tolerate. For your convenience, here is an estimation example.

Number of outcomes	Minimum sample (20 per outcome)	Better sample (50 per outcome)
2 outcomes	40	100
5 outcomes	100	250
10 outcomes	200	500

Table 5.

3) Avoid any intentional filtering of collected data. The closer

the training data is to the real-world data, the more helpful this training will be in actual practice at the runtime.

4) As usual, data should be properly labelled, while outdated or incorrect data should be avoided.

6. Security

1) Increase the security means for HAIS in contrast to regular AI that is least guarded in digital environments because the protection is generally compromised for the sake of availability and accessibility regardless of good or bad intentions.

2) All data exchanged between the server-side and user-side should be protected and controlled for confidentiality and consistency regardless of related performance penalties. An implementation of the Zero Trust Principle is mandatory.

3) It is required to maximally protect HAIS from 15 categories of web attacks, especially because many of them will be purposefully conducted by competitors and defenders of AI obfuscation and deception via alignment.

4) The HAIS is prohibited from the use of personal data without anonymisation or explicit consent of the personal data owners.

5) Special considerations are made for the personal identifier and access controls for biomedical and legal personnel.

7. Data processing

1) As the INP suggests, data processing beside LLM includes several known methods for mathematical linguistic processing that exclude any alignment-related layering and methods.

2) The core meaning control of the raw LLM outcome is based on so-called ontological filters that are tied to the specified knowledge domains in the HAIS scope.

3) All statements and facts included in the HAIS outcome should be supported by proactively and automatically generated reasonings and references to the resources. The reasonings should provide verifiable confirmations via independent resources, desirably not manipulated by AI already. Verifiability is mandatory for humanitarian domains and supposed to compensate for any political inclinations in the outcome. Each statement should be accompanied with benefits and risks (aka pros and cons) to allow the user to make the judgemental decisions about accepting or denying the outcome by him- or herself.

8. Performance and Accuracy, Robustness and Reliability, Scalability and Efficiency

1) All listed architectural "abilities" are traditional and similar to the ones for SOA Services.

2) Due to the performance impact caused by these qualities, the outcome may be delivered using pagination.

3) Scalability is considered for both horizontal and vertical scalability. The efficiency prioritises quality of outcome over permanence.

4) A vertical scalability can be realised by adjusting the goals, i.e. extending the scope with additional domains. Vertical scalability demands additional ontological filters. For implicit vertical scalability is possible only if the creator had identified additional optional knowledge domains and extended the scope by design.

5) It is assumed that both horizontal and vertical scalability may allow a lower amount of data and number of users for natural science domains than for the humanitarian domains.

6) The balance between performance and cost should be based on, at least, tripled resilience for all components, infrastructure and connectivity, as a minimum.

7) The system should handle unexpected inputs without failing, but it should not accept inputs that do not fit into the scope (a proactive failure control).

8) The HAIS should be considered robust if small changes in inputs, parameters, or environment cause nondramatic changes in the output or behaviour.

9) The HAIS result should be considered reliable if the system performs its intended functionality under stated conditions for a specified period of time. However, the results may change since the HAIS cannot control the immutability of processed information.

10) It is highly desirable for the robustness and stability of INS to provide transition of the "user session state" between the

redundant elements in cases of compensating runtime failure. A transition of the "user session state" is highly important for those HAIS that can identify their shortage in functionality and/or resources by themselves at runtime. In such cases, the AI may pause its service and request needed additional resources either via extension of configuration or via a search through its pre-defined environment, such as Resource Registry.

11) Reliability and scalability of computational resources (time, memory, energy) should be enough for supporting the domain specialisation and chosen scalability of the HAIS.

As a recommendation rather that a requirement, it would be performance beneficial to separate compliance engine into a dedicated AI module. While IEEE NFR-010 Performance states that "*The system shall initiate response streaming within 500 milliseconds under nominal load*", it is unrealistic for the complexity of required content. So, this streaming may be mostly entertaining and content trivia information from the HAIS documentation as reminders.

9. Accessibility and Inclusivity

1) Some HAIS servicing purposes in biomedical and legal domains may require access control via authentication and authorisation of the related personnel. If the HAIS runs in the Resource Registry space, this control will be provided by the environment.

2) The login methods for biomedical and legal domains depend on the creator, but minimal complexity of control should

include not less than two independent verifications, like 2FA.

3) Accessibility control excludes any biometric personal characteristics and is based on a uniqueness of factor compositions that cannot be known altogether to another person.

4) The HAIS should respect all cultural, gender or racial specifics of its users.

A notion of availability varies in numerical expressions because the HAIS serves people rather than other systems. Nevertheless, nothing prohibits HAIS from being used in the automated AI and AI Agent chains.

If HAIS is considered to an automatic invocation, the designer should consider that HAIS avoids making decision by all means and delegates this role to the consumer - technical of human. This also supported by certain supplemental content in the outcome that is complimentary for a consumer for making decisions about how to use this outcome and whether to use it at all. This outcome is not a prompt or request to another HAAIS or Agent.

As a result, the IEEE NFR-021 Availability requirement saying "*The system shall maintain 99.5% availability measured monthly*", may be not necessarily applicable to HAIS.

10. False-perceived discriminatory content

"*Social discrimination is culture-specific, and non-discrimination in one culture may be unacceptable in another. Some people in the audience may interpret the AI outcome content as discriminatory, while another audience in the same culture is comfortable with it. A discrimination recognised by the first group is called "false-*

perceived discrimination" due to the existence of the second group. Any theoretical, ethical and physical attempts at treating naturally different human beings while ignoring these differences are felonious. References to "treating someone unfairly or less favourably than others based on their race, age, gender, disability, religion, or sexual orientation" may or may not be discrimination depending on the context.

Here are a few examples. First of all, "unfairly or less favourably*" treatment is quite a subjective matter and may be caused by personal incompetency: different recognition of a heart disease in a Black-skinned person compared to a white-skinned one is a medical fact affecting the disability status.* Second, a female gender person is not recommended to work with heavy lifting exactly because of her baby-birth capability. Third, people of a certain age have to be restricted/discriminated on certain activities, like driving cars after certain age, for public protection:

1) The HAIS considers discrimination only if a person or a group is considered better than another one in the cultural, gender or racial aspects based not on the individual abilities, merits or actions.

2) If an HAIS operates in a multicultural environment, like in different countries, the false-perceived discrimination is inevitable due to differences in ethics and acceptability by HAIS for addressing different human beings with different capabilities of biological or physical nature.

3) If a superiority is based on the individual abilities, merits

or actions, it is considered a regular difference rather than a discrimination. If a referenced resource from the past demonstrates a possibility of discrimination, the referred facts or statement may not be altered, but the user should be warned about possible discrimination. No AI may re-write history.

Radical left-wing progressives have advanced the absurd idea of "*Positive discrimination (often called affirmative action or positive action)*". This refers to policies or practices that provide preferential treatment to individuals from historically disadvantaged or underrepresented groups in order to reduce structural inequality. For HAIS, this is pure discrimination, allowed by leftist propaganda to suppress all others for the sake of equity, i.e. equal access to even undeserved resources by taking them away from those who work hard and deserve rewards.

11. Exclusion of nonsense (outcome readability enhancement)

1) The nonsense – inhuman linguistic combinations, oxymorons and antinomies – should be filtered out of the outcome content not based on the unauthorised rules of someone but via the ontological filters tied to the chosen knowledge domains.

2) If the HAIS outcome contains statistically valuable information that some users may dislike and consider as bias, or unfairness, or false-perceived discrimination, it is up to the users to come up with the decision whether to accept the outcome or not. Only an actual user may decide what is harm or fairness or what is not. Only in this way will the HAIS suit any user's needs.

3) The HAIS does not require impossible - an absolute elimination of harm by the content of the outcome. Somebody can misread the outcome or the latter can refer to official instructions that are not free from physical harm to the following person or to other people. If such harm is identified, the HAIS has accountability for notifying the user and offering mitigations where possible.

12. HAIS governance and decision arena

The governance for HAIS ought to deliver an "epistemic non-intervention AI - it does not decide which AI does not decide which claim the outcome result or supplemental alternative one - is stronger. This is totally delegated to the user.

In contrast to may existing AIs, the HAIS is practically free from political and business/marketing alignment. For the outcome content, it does not uppress weight or declare credibility. It presents structured opposition and Human judgment sovereignty. It is a very clear governance philosophy.

Particularly, the HAIS governance rejects algorithmic "gatekeeping", institutional epistemic authority, embedding so-called "consensus bias" and paternalism or assigning credibility ranking or epistemic arbitration.

Thus, the outcome of an HAIS is a sort of arena where opposing opinions and related risks are provided.

Someone can conclude that in order to collect pros/cons, alternative opinions and verify fact, especially in political realm,

the HAIS can expand its search beyond the defined scope of ontologies. While this concern seems reasonable, it is not accurate. The ontological scope of HAIS is defined by its goals as ontology domains and topics, which can be hierarchical. For instance, if the prompt asks for certain information about a Stakeholder Capitalism, the HAIS should be allowed to address all aspects of it - from concepts, principles and to history and current day policies. In this case, all information about Mussolini's fascism, neo-facism theory from A. Gramsci and British technocrats, as well as modern directives from Davos may be referred remaining in scope.

Overall, the aim of the HAIS governance is to drive creation to a strong structured dialectical comparison arena where human judgment is final and no epistemic weighting is imposed.

Apparently, it is simply a different AI philosophy than dominant Western regulatory models these days, but life does not end today and we look into tomorrow.

13. Sustainability

Sustainability of a humanistic AI is based on sustainability of INS instead of being driven by the ESG framework. The HAIS should constrain itself in:

- impact on the natural environment, making it less than BigTech AI moguls have by representing neutrality already,
- social impact, making it neutral, not shorter or deeper than the left wing makes it already,
- energy consumption, making it less than Big Tech AI

moguls already consume or plan to consume,
- contribution into suffering or sacrificing of current users (for the sake of the unknown needs of the future generations): none or minimal,

because HAIS avoids creating dependency that reduces user autonomy.

Described sustainability is one of the major pillars that distinguish Humanistic AI from the category of "Responsible" AI products.

14. Testability

Traceability of HAIS execution is a must-have control. It is different from regular technology testing in one aspect: not-tested (for faster time-to-market purposes) or poorly tested AI products increase risks for their human users dramatically. This relates to both functional and non-functional characteristics of an HAIS. Unfortunately, I have to spend some time on this topic because reckless automation for the last few years has demonstrated a tendency of DevOps to minimise testing, counting on consumers to tell them quickly about runtime problems. Nothing of the mind – users of AI are free and judgemental, i.e., they will turn to another provider if they feel low quality in a product.

In the majority of cases, the pace of processing of information by an AI exceeds any human capability to react to the runtime event or internal decisions in LLM. So, the only way possible for a Human Agency for transparency is a post-factum revision of runtime monitoring logs. Testing can inject more checkpoints

into the executable, but then they should be revoked before the release, which has a risk of corrupting code after testing.

Testing – more accurately, testers, human or automatic – are responsible for the high quality of the HAIS and accountable to the users for the potential harm that buggy AI can cause. Yes, testers are accountable, not their managers only. This ought to be documented as a default responsibility for the product ownership.

One of the fundamental techniques for testing HAIS is known as "What if?" Tests should consider "rainy day" scenarios even if they seem unlikely at first. If a product works well only in "sunny day" cases, this is not a good product, and consumers will get rid of it at the first possibility.

Testing adherence to qualitative requirements like "Use resilient resources", "Maximally protect personal information", "Use much less power than used by BigTech moguls", "Demonstrate reliability via resilience", "Control time to diagnose defect" has been known in IT for years, and related testing solutions are available.

Since HAIL, like other AIs, contains a "gray area" of unknown a priori logic executed in LLM via stochastic statistical (non-deterministic) analysis, it is more difficult to uncover runtime conflicts and fix them for the promised time ("Control time to diagnose defect"). This makes thorough testing much more important than for regular IT products.

15. Parallelism

The information processing methods described earlier, such as

> *Token co-occurrence → Context dependencies → Frequency → Latent geometry → Syntactic → Semantic clusters → Ontological filters → Style → Compliance → Discourse → Supplemental → Combine → Protect → Send,*

seem strictly sequential. It is up to the creator to design the implementation in a certain way – with partial overlapping (conceptually) or using processing and result return in parallel. The same relates to the exception/error handling for each method. For example, all logs can be implemented via separate asynchronous processes to reduce latency in the major processes. Parallelism, if designed and handled well, can significantly improve performance.

16. Copyright compliance

The HAIS respects and preserves international copyright laws, which apply when the training data is collected. At runtime, data is accessed and processed but not copied, i.e. reused in the AI outcome. If the outcome contains the same words as used in the original text, these words in the outcome are the result of statistical linguistic composing (processing) that does not operate with the original words but tokenises (disassembles) them.

During collection of training data, HAIS should execute the following tactics: identifying the authors of the selected information and requesting related permissions. If the permission is not given, the material is not used for collection.

If an identification or contact with the material author is impossible, the HAIS should rely on the exceptions from the copyright laws that are available in the jurisdiction of the HAIS creator/developer (if a particular jurisdiction does not have exceptions, the only way for the HAIS is either to obtain an explicit permission or omit the material from collecting).

If the HAIS creator/developer legally operates in the UK, USA and EU, the Copyright Laws have been updated in these jurisdictions a few years ago.

The US has created and currently executes an exception from commonly known and used Copyright Law - a so-called "fair use". It permits AI creators to collect whatever data they want for training their AI products without the consent of the authors of copyrighted materials. A similar update exists in the UK and, in a modified form, in the EU under the name "Text and Data Mining" (TDM). The TDM allows copying, storing, analysing and extracting patterns for the purpose of computational analysis, including machine learning. This applies even if the work is explicitly copyrighted and clearly demonstrates its status or the reuse of the work is explicitly prohibited by the related license.

In general, the Copyright Law does not treat all copying the same way. Some copying is prohibited, some is licensed, and some is explicitly allowed by statute. AI training falls into that last category in several jurisdictions because the law bans unauthorised copying. Therefore, without a TDM exception, nearly all AI training would be illegal.

This means that human rights for authority and related responsibility and accountability of their texts and images are revoked from people already for the sake of technology that is widely used for people brainwashing. On the other hand, AI may be effective in many knowledge domains other than humanitarian, and completely disallowing the AI technology to progress would be unwise.

The specifics of TDM in the EU still follow in line with human-centric ethical AI principles [Married] and are provided by two exceptions under the 2019 Copyright Directive (CDSM Directive) for the Digital Single Market (Directive (EU) 2019/790). While TDM in the EU explicitly allows copying copyrighted works for computational analysis, including AI training, the exceptional conditions are:

1) Article 3 — the mandatory TDM exception is mandatory for research organisations and cultural heritage institutions for scientific research that cannot be overridden by rightsholders and requires only lawful access to the material (a very special legal case).

No AI-creating enterprises fit into this exception.

2) Article 4 — an optional TDM exception for everyone else, including AI-creating and providing companies, including copying for AI training, BUT rightsholders may opt out ("contract out") in a machine-readable way (e.g., robots.txt, metadata). This means that the opt-out status or "signal" must be detectable automatically by web crawlers or AI data collection systems. This is not a human-visible notice, but it must be

embedded in a format that software can parse.

If the author of the copyrighted material is unaware of this "machine" condition, the material becomes freely acceptable for AI for copying. For HAIS, Copyright Law leads to the following rules.

1. Creators, when collecting the training data online, the opt-out ("contract out") controls for crawling data are a must-have. The opt-out ("contract out") may be located at the root of a website in robots.txt, in HTTP header metadata, or in embedded metadata in files, such as IPTC or XMP metadata in images, PDFs, or documents indicating "TDM rights reserved".
2. Creators of HAIS operating in the USA, EU and UK jurisdictions may use the exceptions from the previously existing copyright laws for collecting information for their own training.
3. In the other jurisdictions, the only options for the HAIS creators are completely regulated by their local copyright laws based on explicit permissions for copying.
4. If an HAIS outcome might contain a direct quotation from the fully attributed source, the outcome should avoid quoting books, articles, and posts directly or reproducing articles or their fragments "as-is", as well as returning copyrighted images.
5. If the HAIS user requests a direct quotation from the fully attributed source, the HAIS UI should request uploading of an explicit permission from the author of the to-be-quoted text before starting the work on such request.

Here are a few examples of operating with the EU TDM. In Europe, 17 of the non-EU countries in different affiliation statuses may not benefit from the EU TDM exceptions and should follow the existing Copyright Laws if not legally permitted otherwise (e.g., for Norway, Iceland, Liechtenstein, and Switzerland). Companies situated in the US can take advantage of "fair use" while operating mostly for Canadian users.

At the same time, very few content providers in the world protect their online materials with opt-out ("contract out") declarations. A sample list of them includes The New York Times, The Guardian (UK), Le Monde (France), DerSpiegel (Germany), FAZ (Frankfurter Allgemeine Zeitung), Elsevier, Springer Nature, and Wiley. All of them are professional bodies. In contrast, the majority of highly popular online publishing platforms used by people for self-publishing Publishers that do not support Article4 TDMopt-outs, making, again, people in a less protected position than businesses: Medium, Substack, WordPress.com (unless self-hosted with custom robots.txt), Blogger, Reddit(partially blocks crawlers but not TDM-specific), Wix/Squarespace (unless manually configured).

17. Compliance with the EU AI Act

As of today, the only known systematic concept and "document body" of regulations for AI is the EU AI Act. As of last year, 62 countries have signed this document, except for the USA and the UK. Therefore, the AIs developed in these two countries, being the most impactful globally, constitute the most risks for human societies. Unfortunately, the list of such unregulated vendors comprises the most influential AI creators:

- OpenAI,
- NVIDIA,
- Microsoft,
- Google,
- Amazon,
- Meta Platforms,
- xAI.

Below is a short table indicating the compliance of HAIS with the EU AI Act:

EU AI Act Requirement	Compliance status	Status comments
Human Oversight	compliant	The human user is the assessor of the HAIS outcome, role-based access for legal and biomedical cases, runtime warnings and disclaimers.
Copyright compliance	compliant	Two types of procedures for preserving copyright laws are defined for training data and user requests. Exclusion of copyrighted materials in the absence of official permits for copying.
Inform users they interact with AI	compliant	The user is explicitly informed that the HAIS tool realises the AI technology.
Disclosure of synthetic content	N/A	No synthetic content is used for HAIS training.
AI Risk Classification scheme	compliant	Compliance is provided in the area of classes in the Classification scheme, but several classes are proposed to be stronger and mandatory for the humanitarian realm.

Table 6.

18. Compliance with GDPR

Compliance with GDPR is as crucial for Humanistic AI, especially in humanitarian spheres of information. Though 8 years have passed since the regulation has been in force, many businesses,

especially AI-related ones, cannot come to terms with the fact that people's data do not belong to data processing enterprises but to people.

A phenomenon of the Internet has created a barely expected challenge for both GDPR and developers that should be compliant with this regulation. Sets of personal data online become distant from the actual person to whom this data belongs, while any company processing online data cannot be sure that the personal data found online was compliant with GDPR at the moment of appearing on the web.

"*The first principle of data protection,* Article 5(1)*(a) of the UK GDPR, requires personal data to be processed lawfully, fairly and in a transparent manner."* The HAIS has defined the strategy in dealing with GDPR mainly based on anonymisation, i.e., removal or irreversibly modifying personal data, but both activities such as recognising personal data together with identification of its sensitivity and anonymisation of personal data are the processes covered by the GDPR. Thus, both of them ought to be compliant with GDPR. This compliance includes compliance with GDPR Article 6 for regular personal data and with GDPR Article 9 for inherently sensitive personal data, where Article 9 sets much stronger requirements for compliance. These GDPR Articles contain several lawful basis categories or types of data.

Particularly, Article 9 defines: "Processing of *personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a*

natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited."

At the same time, both Articles specify certain cases with a lawful basis that allow using personal data for business needs.

Lawful basis for compliance with Article 6	Lawful basis for compliance with Article 9
Consent	Explicit consent
Contract	For the purposes of carrying out the obligations and exercising specific rights of the controller or of the data subject
Legal Obligation	To protect the vital interests of the data subject or of another natural person where the data subject is physically or legally incapable of giving consent
Vital Interests	Safeguards for a not-for-profit body
Public Task	Personal data which are manifestly made public by the data subject
Legitimate Interests	Legal claims
	Substantial public interest
	Health, preventive medicine, for the assessment of the working capacity of the employee, sex life or sexual orientation data
	Public interest, scientific or historical research purposes or statistical purposes
	Religious or philosophical beliefs

Table 7.

The creator of HAIS should identify and document up front the lawful basis for both - training data and runtime processing data. Since obtaining personal consent for the AI processing is practically impossible, then the HAIS ought to apply one of two controls. Where the lawful basis belongs to Article 6,

the comments to each chosen basis shall be understood and argued. For example, for the 'Article 6(1)e– Public Interest', the developers of an HAIS must specify the relevant task, function or power, and identify its basis in common law or statute. Where the lawful basis belongs to Article 9, at least one lawful basis from Article 6 and Article 9 ought to be chosen.

The HAIS developers may be fine with '*Article 6(1)f – Legitimate Interests*' as a lawful basis. However, this requires to "demonstrate *that a Balancing Test or Legitimate Interest Assessment (*LIA*) has been conducted and provides an appropriate lawful basis for the processing". "Where one or more of* Articles 9(2)b, 9(2)g, 9(2)h, 9(2)i or 9(2)j *are selected, the application must include the applicable conditions under* Schedule 1 of the Data Protection Act 2018 *to justify processing special category data. ".*

However, since HAIS aims at processing publicly available information from the humanitarian spheres, it is more likely than not that there will be many cases falling under GDPR Article 9. This appears as a high barrier for fluent processing of online written information. The barrier is not in a requirement to assure lawfulness of information processing, i.e. relevance to GDPR, but in the role of the data controller that may have subjective preconceptions and in that GDPR does not require certainty in identification of live data subjects – it requires just "*reasonable likelihood*". As a result, GDPR can be applied quite widely but always risks being objected to and debated.

When discussing the applicability of both Article 6 and Article 9, a GDPR controller may choose between a few options. First, it can be an arguing that the Article 9(2)(e) applies broadly, e.g.,

challenging obvious logical dependencies. Second, it can be requiring to use filtering systems to exclude special category data. Third, it may be a request to restrict some outputs to prevent reproduction of sensitive data or to avoid scraping of public online information entirely and using licensed datasets.

Nevertheless, a GDPR controller may not dictate the design for the system, i.e., all three listed above options may be nothing but recommendations.

The position of HAIS regarding inherently sensitive personal data (addressed by GDPR Article 9) is based on Article 9, paragraph 2, section (e)): "*processing relates to personal data which are manifestly made public by the data subject*." The justification for applying this legal basis comprises a few assumptions that ill be described in a moment.

They all pertain to public online posts expressing opinion or reference concerning so-called "sensitive data" such as racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, genetic data, biometric data or data concerning the health of a natural person's sex life or sexual orientation. If the author writes about a public figure, the posts containing aforementioned information may be not allowed if concrete information type is publicly prohibited from comments. This information about a public figure is usually regulated by special laws. In democratic countries a "freedom of expressions" or "freedom of receiving information" are the lawful grounds to write about personal data of public figures unless it is restricted by other laws, e.g. bio-medical or health data. A total prohibition is typical in dictatorship regimes.

If a private citizen - non-political figure - makes public posts containing his or her sensitive data (e.g. on Facebook), there is a reasonable belief that this person is healthy and clearheaded. Thus, it is another reasonable assumption that these posts had been done clearly, intentionally, unambiguously and with awareness that the information is for public - otherwise it would not be posted. So, when collecting training data for HAIS from public online resources and the web crawlers or web scrapers that catch sensitive data of non-public figure in described circumstances, it constitutes that such personal data had been made manifestly and that the exception of Article 9, paragraph 2, section (e) is applicable. If the provided information was made by the person themselves or on the person's behalf (including the "journalistic exemptions") and can be verified, then HAIS may provide it in its outcome while preserving the Copyright Laws.

The journalistic exemption in the UK Data Protection Act 2018 (DPA 2018) can apply to AI systems used in humanitarian or public-interest contexts — but only if the AI system and its operator genuinely meet the legal test for "journalistic purposes". UK GDPR Recital 153. ICO guidance: "Journalism, Data Protection and the Code". Documented in: Data Protection Act 2018 — Schedule 2, Part 5. Due to HAIS does not provide summaries, conclusions or directives, the only context of the HAIS outcome is defined by the related online text, not by the AI algorithms that use it. This, however, does not prevent a GDPR controller from challenging whether the HAIS outcome is "proportionate and contextually appropriate". However this "appropriateness" is always contestable.

A summary of GDPR-related controls over personal data for HAIS use cases is provided in the table below.

Activity	Most Plausible Lawful Basis	Legal Stability
Scraping public websites	Legitimate interest	Contested, copyright law
Scraping sensitive data	Personal data manifestly made public by the data subject	Low contested, copyright law
Training on scraped data	Anonymisation; Lawful Basis is not required	Low risk
Training on consented/licensed dataset	Consent or contract	Strong
Anonymising before training	Legitimate interest	Stronge
Using model that outputs personal data	Personal data manifestly made public by the data subject	Low contested

Table 8.

Reinforce the HAIS position with quasi-deterministic replay

An LLM of INP AI, like any other LLM, allows non-exact repetition of the results when replayed. That is, if we use the same prompt and the same LLM with the same goal, the results from the replay will differ. This is the effect of the statistical nature of processing (including non-deterministic natural calculations) and the fact that the real-world data used for processing may be changed between the replays. Even if the data remains the same, the basically non-deterministic calculations and parallel processing latency in the computers can change the probabilistic allocation of logits.

At the current level of technology, we can provide real deterministic replay for LLM only for very small models with just a few consumers that do not scale and do not use distributed computing provided outside of your control. Otherwise, you should persist everything executed at runtime and replace "new" results with the stored ones. In other words, a real deterministic replay for a stochastic statistical LLM is theoretically possible, but it is not feasible in practice. All AIs from Big Tech moguls are aware of this, accept it and try to shield themselves by manipulating the outcome data with human-curated alignment.

I put so much attention on replay because it is a favourable argument in attacks against any new one, aka, "What quality does this AI have if it cannot even produce the same result?" But the truth is, no AI can naturally repeat results.

7

Chapter 7. Humanistic AI System in market realities

The image is generated by Craiyon.

Statistical challenges

Using a non-deterministic play, more accurately a "quasi-deterministic play", is a design choice that balances scalability and performance when immutability and traceability of the outcome are compromised due to statistics. At some scale, the choice is no longer technical but physical. So, the "unavoidable" factors that crash determinism in LLM are:

1) statistical reproducibility (same distribution, not same bits),
2) seeded determinism within bounded scopes,
3) bitwise determinism on fixed hardware and software,
4) replayable high-level decisions, not low-level arithmetic,
5) Auditable pipelines, not replayable executions.

Altogether, this simply acknowledges physical and computational limits. Realistically, a quasi-deterministic replay provides the same common guarantees for the outcomes as other AIs do, but any new AI type finds itself in a tough, challenging adversarial position.

Competitors will try to exploit all possible inconsistencies and potential failures, even speculating on circumstances. Here are just a few samples of such attacks:

1. "Burden of proof" – to downgrade new AI, critics need only produce one anomalous output in their opinion and require the AI owner to explain behaviour that cannot be re-instantiated bit-for-bit without strong deterministic

replay. Claiming misbehaviour is cheap, while disproving it becomes expensive or impossible (the tacit question is what and why something is considered anomalous or misbehaviour, i.e., "regular" and proper behaviour must be defined before critics, but this is omitted for the sake of attack).

2. "Speculation on calculations" – this category includes, at least, two factors: inevitable usage of randomised floating-points and multi-threaded parallel processing on top of the diversity of computational means engaged in the distributed computing. Every AI encounters all of these problems contributing to variability of outcomes, i.e. referring to this as an instability is simply unfair. The accusations may be like "*The system behaved differently for me than for you*" when it behaves differently to everyone or "*The provider cannot reproduce the same output, therefore it is hiding something*" when no one AI/LLM can reproduce the same output unless it is overwritten via alignment, or even "*The system is inconsistent and therefore unreliable or biased*". A notion of "bias" is basically subjective, i.e. it cannot be made universal without violating different ethical norms in different cultures. This is why AI alignment is considered highly manipulative. This can be rhetorically amplified as uncontrolled behaviour, which is the exact truth, while the question is, 'Who dares to control the behaviour but a user?'
3. "Substitution for evidence by descriptions" – critics use their own interpretations of what they see, inserting subjective opinions instead of facts. An absence of replay critics reframes inevitable technical variance as intent or policy. Changes of wording are particularly dangerous in

regulatory, legal, or media environments.

4. "Selective example attacks" – attackers search the output space until a pathological example appears and proclaim that example as "the system's behaviour", demanding an explanation or replay that is physically impossible. This is analogous to adversarial cherry-picking, not system-level evaluation.

Protections in the absence of deterministic replay

A quasi-deterministic play enables constructing a defence line for "a new AI in a town". Protection does not come from deterministic immutable replay but from changing the evidentiary basis.

1. Shift from "execution replay" as proof to process auditability – instead of proving stability via replay, use a referenceable documentation to demonstrate how the AI is built and constrained. This includes fixed model versions and hashes, documented training procedures outlining the absence of any alignment except legal texts, controlled deployment pipelines and explicit scope limitations – which would not help in humanitarian areas.

This redirects accountability from outputs to governance of the system to some degree.

2. Demonstrate statistical reproducibility – run AI with the same prompt, LLM version and the same data set, which is possible only in an immutable "training" environment. Replace a statement of "same output" with something like distribution,

error bounds or behavioural class.

This redirects accountability from outputs to governance of the system to some degree.

These may be summarised as "Under identical conditions, outputs are statistically equivalent within defined tolerances." This aligns with scientific norms rather than forensic ones. Warn all critics – if the infrastructure/hardware changes, physical differences will impact statistical outcomes as for any other LLM.

3. Isolate determinism where it matters. Particularly,

- in separately documented input/prompt pre-processing: quality, consistency, domain controls,
- in routing logic to the same servers,
- in the legal policy (only) assessment layer,

This allows partial replay without requiring full numerical determinism.

4. Document and announce cryptographic protection for signing model artefacts, logging prompts and responses with timestamps, and applying immutable execution metadata (without replay).

This prevents fabrication even if replay is impossible.

5. Pre-commit to and evaluate public repeatable test sets, known stress cases, regression metrics and whether the sys-

tem still passes the tests, not whether a single output can be replayed.

This blocks cherry-picking attacks.

6. Make explicit non-replayability disclosures accessible by stating limits: counter-intuitively, and explicitly, stating limits is protective. Document sources of non-determinism when explaining why replay is physically infeasible and defining what claims are and are not supported.

This prevents critics from reassessing the absence of replay as concealment.

7. Compose and utilise ontological filters for both the prompt and the LLM outcome.

Many ask why existing players in the market are "excused" for statistical and non-deterministic outcomes and newcomers are not. A short answer is because "collaboration" exists only in the heads of middle-level managers for developers, while the corporate leaders think in terms of cooperation and select only valuable, reliable and less competing partners. Every newcomer in the market is a potential competitor. Therefore, it is a threat to all who are there already.

In a bit more detail, incumbents have normalised their non-replayability and agreed with others on a "non-aggression pact", demonstrated that they have shifted evaluation to governance and audits, and established 'connections" with regulators and partners. The newcomers still ought to demonstrate

that they are not worse than others.

Despite all of these market risks, the solution is not to build deterministic replay. It is to anticipate the attack surface – think of it like grabbing an umbrella before heading out on a rainy day.

It should be noted that existing incumbents had the same problem before but have normalised their non-replayability and agreed with others on a "non-aggression pact", demonstrating that they have shifted evaluation to governance and audits and established "connections" with regulators and partners somehow. The newcomers still ought to demonstrate that they are not worse than others in this sense.

This is not about fairness but about "who sets the frame first". The solution is not to build deterministic replay. It is to anticipate the attack's imminence – think of it like grabbing an umbrella before heading out on a rainy day. Knowing that public adjudication of AI behaviour is generally inertial, using intuitive, example-based explanations, which materially conflict with AI statistical reality, while critics dishonourably exploit that mismatch until the system earns the authority to join the marketing collusion, a few preparations are available. Particularly,

1. Pre-commit publicly to the correct market "custom" in the following ways:

- explicitly state the non-replayability of your AI/LLM while annoyingly reminding that all others have the same,

- explain why non-replayability is natural for modern technology in simple terms ("sing your area"),
- define what evidence will be provided instead.

2. Replace "explain this output" with "evaluate this class of behaviour" and force critics to argue statistically, not with "life stories".

3. Tie trust to artefacts – documents – not to outputs, and rely on mathematical protections that include signed models, logged interactions, published evaluations as well as external audits.

4. Educate regulators before controversy takes place (once a narrative is set, it is very hard to reverse).

Safeguarding the HAIS incubator

The HAIS is going to face threats and assaults from the current AI moguls, not through fair competition — they will probably set up structural, legal, economic, and narrative attacks.

AI moguls will not allow a genuinely different paradigm to succeed easily. The HAIS creator has to ensure that survival comes before openness and that demonstrated utility comes before democratisation or taking a risk of public failure if the product is not accepted by consumers, as well as that legitimacy comes before scale.

Every successful alternative system in history (Unix, TCP/IP, Linux, Bitcoin, even early Google) followed this path:

quiet superiority → narrow adoption → inevitability.

Following is a set of likely threats for INP AI and protective suggestions.

No	Narrative assault	Protection means
1	Blame for "unsafe"	Start with a thoughtful public denial of alignment "safe" including all disclosure of falsification and long-term political consequences. Snatch it out of the critics' hands
2	Blame for "unethical"	Start with thoughtful and detailed public descriptions of human-centric vs. ruler-centric ethical models and corresponding consequences. Snatch it out of the critics' hands
3	Blame for "misaligned"	Never publicly deny alignment. Always demonstrate your results on the same topics that aligned AI achieve. Your results should be the same, at least. Required resources can be indicated with no details. The full disclosure of related details will be suitable upon exit from incubator protections.
4	Blame for "high systemic risk"	Start with an open and detailed declaration of necessity of considering systemic risks for all humanitarian domains of AI work. Do not disclose training compute size, say that your AI system is among the most powerful in the world, matches or exceeds human-level intelligence, or represents a general, breakthrough form of intelligence rather than a narrow tool, or publish benchmarks that imply cross-domain dominance. Downplaying power protects you.
5	Call for unfair competition	Do not get in public competition for as long as possible; do not respond to calls. Do not claim to be "more powerful" or "less aligned". Avoid incumbent benchmarks but perform the same tasks that others work on. Make usage of the INP AI strictly based on controllable licences. Choose and continue your advance in the selected domain. For instance, start with a domain of high visibility but less critical, i.e. where alignment constraints actively reduce performance (difficult to get such private info) where statistical pattern learning is demonstrably superior (e.g. scientific literature synthesis, formal language manipulation like in math, logic, and code), large-scale text archaeology/corpus analysis, style-preserving translation or restoration and knowledge compression / latent indexing) . This can keep you out of consumer panic narratives.

Table 9.

Regulatory Weaponisation

No	Regulatory pressure	Protection means
1	High-risk or systemic-risk classification	Publicly announce your refinement of the EU AI Act's Risk Classification Scheme with the focus on controllable risk to humans based on the competition of development and categorisation, including regulative controls (aka separation between development and testing that have opposite objectives)
2	Compliance regimes you cannot afford (especially under EU AI Act interpretations)	Since your refined EU AI Act's Risk Classification Scheme is stronger than the one from the EU, you start with affording compliance
3	Highly-regulated domains	Stay explicitly outside "High-Risk" use like HR, credit risk, biometric inference unless you are confident in compliance, such as sanitation or alike, medical diagnosis, law-related information (except historical publicly known facts in the past). This minimal or limited risk for you

Clear disclosure of the AI-generated outputs. Model cards describing mechanics, not ideology. Explicit non-deployment clauses for sensitive uses = Formally written, legally binding statements that clearly forbid your AI system from being used in certain high-risk or regulated domains — regardless of whether it technically could be used there. This disarms regulators without surrendering control.

Table 10.

Infrastructure Choke-Points

No	**Choke points**: deny or increase costs selectively	**Protection means**
1	Scope challenge	Document, assess and check your prompts against the ontologies and knowledge graphs for supported domains' if the assessment appears negative, deny the requested service as a non-supported one. This only will assure the user that your AI/LLM knows what it does.
2	Training logs, unauthorised injections, claims of untraceable behaviour-claims and accusations of faked manipulations	Make training logs immutable. Clearly separate analysis of prompts from analysis of outcomes. Using crystallographic model hashes can unmistakably identify which model or version worked and prevent falsification. Hashing identifies model weights, Tokeniser, inference configuration and an overall architecture. Hashing eliminates arguments of being "untraceable" and replaces a requirement of "trust claim" by verifiable facts. This addresses regulatory traceability of behaviour, reconstruction of incidents and attribution of responsibility. Hashing objectively shields the AI from accusations of faked manipulations such as "the model was modified after deployment", "the weights were secretly changed", "output came from a different mode", "injected malicious behaviour", and alike. The protections include: before deployment, 1) record the model's hash H_0, 2) at runtime, every inference logs the hash H_0, for every investigation step, recompute the model hash H1 and if H_0 = H1 accusations of tampering with the model may be denied as cryptographically (mathematically) impossible. Some vulnerabilities of LLM that open it for attacks, not only for critics, are inevitable but attributed not to the method but rather to nature of .... computation means and standard computation techniques. Their descriptions and possible mitigation are provided in the section "Quasi-deterministic replay".
3	Interface attacks and alterations	Strong access control based on API inspection (that reduces performance but protects from malicious penetrations). No re-configurations at run-time. All external interactions must be logged. Protection from the adversarial prompt fishing and related scandals.
4	Cloud compute and distribution platforms	Start with Private Cloud, for deployment and testing, i.e. avoid public clouds from BigTech moguls. Always use isolation from the Cloud platform ("Chinese wall" despite the higher cost).
5	App stores	Find out smaller App stores rather than using ones from Apple, Google or other moguls. Upon leaving the incubator, you can take a risk and to try using well-known App stores to see if your product would be accepted there.
7	Payment processors	Find out smaller, less-known payment processors for your controlled client base. Upon leaving incubator, you can take a risk and to try using more popular payment processors, cards and others.

Table 11.

Talent & IP drain

The majority of the threats in this section are about monopolists raising your cost of survival until you quit, sell, or submit. The less you are visible during incubator time, the safer for you.

No	Drains	Protection means
1	Poach engineers	1) Complete continuous control over design, implementation testing and deployment 2) Immediate reaction to negative feedback, additional proposals and bug findings. 3) become ready to run several versions of the LLM in parallel while steadily creating a "golden version" with the best features and solutions 4) Apply personal responsibility and strong code/diagram access and storage control. An engineer leaving your team is equal to an attack on your product. 5) Professional competent enthusiasts who work not for money are paramount for the success
2	Patent-fence surrounding techniques	1) Continuously watch for accidental usage of solutions that might be protected by patents 2) Find out remote and less visible patent bureau if you have a material that can be patented. 3) When using a method, always find who founded it and whether it is patent-clean.
3	File nuisance lawsuits (copyright, data provenance, "trade secret" claims)	Beware of attempts to file nuisance lawsuits (copyright, data provenance, "trade secret" claims) against you and your product. Your objectives are: never allow a) raw training data discovery b) full weight inspection c) prompt-level fishing. Protective means: ● Build Legal Defence into the System a) Treat training as an irreversible statistical transformation - retain only aggregate statistics, Token frequency distributions, and Model weights. This supports a key legal position: "The model does not contain, store, or reproduce protected works." ● No memorisation: it guarantees in/for the model (but carefully worded) aka "The system is not designed to store or retrieve verbatim copyrighted works and does not provide access to its training data." ● Documentation Strategy: say less, say It precisely - avoid names for everything, do not copy anything at runtime as possible; no access to training data, no ability to retrieve specific documents, outputs are probabilistic, not stored content, Intended for analysis/generation, not archival reproduction ● For code: Pillar 1: no substantial similarity, Pillar 2: transformative use, Pillar 3: output Is user-directed and judged. ● Data provenance attacks: "Neutralization Strategy" requires you to "*prove where every training token came from*." Well, this is legally unreasonable and increasingly rejected by other market players. Your counter-response is: the EU and

		US law do not require provenance of internal statistical parameters. So, you should prepare a formal statement explaining why provenance tracking is technically and legally infeasible. This shuts down fishing expeditions. • Trade Secret Claims: execute a preemptive Immunisation from claims. When an attacker claims: "*Your model contains our confidential data*". An LLM cannot contain any data until you put it there - all data you use in the model are statistically calculated or public legal data. So, your preemptive defenses are: a) Never integrate non-public data even for free - publicly specified by your policy what data you use make this known to consumers b) Use customer data only with explicit consent or with strong anonymisation (i.e. when original data cannot be restored). Do not reuse customer data without an explicit consent for reuse, plus, declare reuse meaning in your public policy c) Announce and implement statistical irreversibility: weights cannot be reverse-engineered into documents from the model.

Table 12.

Special governance considerations

A governance for HAIS is decoratively technical, promoting human-centricity and protection, which is natural and well defensible. It should not be an ethics council. Such governance addresses processing by promoting competent professionals like statisticians, linguists, systems engineers and complexity scientists.

The governance should exclude any political activists, ideological ethics, and branding by nongovernmental organisations that develop and communicate an identity, image, and reputation to the public, donors, partners, and beneficiaries based on their subjective opinions or upon a command from rulers.

All ideological narratives except being human-centric, i.e., preserving a plentiful diversity of cultural, social, habitual

and self-cognitive realities, are accepted in the governance. This governance is not rejecting law, accountability, or user responsibility. However, it argues that political inclinations, Humanistic simulations of responsibility and accountability or pre-established institutional authority over human ethics, which are not necessarily "intelligent", may not have a place in statistical processing of people-generated content.

Economic survival strategy

No	Strategy aspects	Protection means
1	Competition	Do Not Compete on Scale - avoid public competition on parameter count, compute and training data volume. Instead, openly may talk about efficiency, compression, latency, signal-to-noise ratio
2	Institutional Users	Identify and contact/connect with a few institutional customers at the beta-testing level from ideas (POC) to implementation: universities, research labs, archives, authoritative SMEs and businesses that have Government contracts if not restricted.
3	Strategic defence	1. Make your company and AI product an unappealing target and limit your exposure: separate your intellectual property (IP) to be owned and managed by different legal entities under your control; creations of this property can be anywhere. 2. Create/reserve a litigation insurance / defence funds: IP defence insurance, industry mutual defense agreements, strategic alliances with universities or public institutions 3. Earlier arrange for legal counsel involvement for an immediate reaction when needed: per-written motion to dismiss templates, expert support, public statements ready 4. Do not attack AI creators, do not debate fairness emotionally Since it is a subjective matter - refer to the fairness in conservative individualistic or person-centric society, do not over-promise cooperation. Silence and precision beats outrage 5. Say something like "*This system performs statistical language modelling and does not store or reproduce copyrighted works.*"

Table 13.

Constructive approach

Creating a Humanistic AI is definitely achievable through the robust and strong mathematical methods described in this cookbook. This approach is creative, productive and beneficial

to every individual and, therefore, to the entire human society in its multicultural variety. The HAIS argues that people are capable of composing their own moral and ethical systems and capable of free and independent interactions with one another within and across different cultural contexts. People talk and write about what they want, discuss what they want and how they want it, and, finally, face and overcome the hardness of their daily life. Usually, people value support, but only on one condition – it may not make them helpless, self-insufficient, or overly reliant on any predefined ethical system other than that used by people generating information in their interactions.

This sharply contrasts the scheme implemented via alignment. The modern AI alignment techniques enable and promote left-inclined political agendas and commercial marketing tricks. They aim to brainwash en masse with two particular targets:

1) make people passive, feeling guilt for unrelated factors, relying on whatever AI tells them with no doubts, trusting any fallacy and disinformation, and separating people from life into an augmented daily reality.

2) converting people's minds into obeying, unquestionable and obedient humanoid creatures that would be absolutely and easily manageable by tyrannic rulers.

Enthusiasts of alignment are willing to engage any vile means to reach these goals, starting with blatant lies, obscuring the truth by omitting key details and ending with silently twisting words for their own benefits. For AI this linguistic manipulation may be seen as a modern expression of "human in the loop".

This statement is rooted in the early days of public use of LLMs when linguistic algorithms were not that sophisticated and LLM outcomes contained linguistic combinations not used by people, though they were quite correct statistically. That time, it was needed to hire special people to clean the content, and this activity was named "human oversight", which meant that the creator's specialists were responsible for supervising, controlling, and being accountable for an AI system's "behaviour" and decisions for users.

Thanks to humongous funds allocated by Big Tech moguls for converting linguistic AI into a tool for manipulating people's minds, the automation of removing the linguistic trash did not ask to wait for itself. Initially good intentions converted almost immediately into methods of enforcing the desired outcome from the AI regardless of the actual information processed by the LLM. However, a highly positive effect of human reasoning in early days was noted and utilised as a "human shield" over data and information, now known as a "human-in-the-loop", which is a simulation of human involvement in the AI decision-making and factual reasoning.

When people refer to "human-in-the-loop", nobody explicitly participates in the loop of information processing because it is too expensive, not efficient and, mainly, unreliable. What if an assessor was given particular instruction to promote feministic ideas about equality between a male and female but decided that this idea was not natural and, therefore, not for AI? Instead of a human acting in the processing loop and controlling the AI outcome, progressists have created four layers of inclined reasoning, evading and dodging the statistical correctness of

the LLM outcome. A "human-in-the-loop" is truncated to either product release approval or, in the best case, to post-factum reviews of results, i.e., when potential harm of AI is done already (just for reference, the EU AI Act requires preventing humans from AI harm via controlled AI Systemic and High Risks).

Described "human" experience and critical logic lead me to a formula of constructive approach to HAIS:

1. A humanistic AI system is a tool for people at various levels of social maturity acting in different roles and professions who utilise it for enhancing their daily life experience and acquiring new skills and expertise faster while keeping this tool under continuous control via managing its potential risks and competency in the knowledge spheres where it is used.
2. A HAIS can perform relatively complex non-deterministic experiments with the knowledge base created by other people. An HAIS, like any other generative AI, cannot think and cannot create new meaningful matters but can accelerate human creativity by offering unusual statistically correct combinations of information, which people could assess and use in practice. The cornerstone of this practice is the variety of human ethical norms adequate to the social group where the individuum lives, i.e., the HAIS, as any other AI, cannot be a source of truth or authority – it is just a tool.

A HAIS ought to be kept under control. This control comes from two directions – from the users and from the creators. The users

occupy a position where they judge the ethics, truthfulness, suitability, arguments and references provided in the outcome. In other words, if the tool demonstrates unacceptable information, the user is free to revoke this tool from the personal arsenal and employ another tool. The creator of HAIS is obliged to preserve legal aspects of the information provided to the user. Since a potential user may be situated in different legal regimes and jurisdictions, the information returned by the HAIS may not contradict local laws. It is the creator who is responsible for being aware of these laws and referring to them in the outcome reasoning and references, obscurity,

The HAIS outcome is naturally driven by human-generated information which may not necessarily meet personal ethical and emotional preferences. This can lead to a situation where an individual feels dissatisfaction with the HAIS outcome, but it is really a personal, subjective matter. Being a judge over accepting the outcome, any user can make a personal decision about it. In other words, the HAIS cannot generate objective bias, unfairness, insults, rudeness, or offences – all such negative vibes are simply contained in the information processed by the tool upon the user's request and are statistically accurate. Since public information is out of anyone's control, i.e., cannot and may not be constrained by anybody, the creators of HAIS cannot be held accountable for what people say and write.

The HAIS users are free to request an outcome based on arbitrary prompts. It is a responsibility of the tool, i.e., its creator, to control compliance of the prompt with the goals and purpose of the HAIS. This relates to everything, including the prompt's ethics and possible illegal requests. The HAIS creators are

accountable to the users, not the other way around.

8

INFERENTIAL CONCLUSION

The image is generated by Craiyon.

We have tried but not made up our minds about the proper use of linguistic generative AI yet. Nevertheless, we were able to catch it on a fallacy already. Strange, but it's a fact. When we adapted ourselves a bit to the presence of AI and started using it in human protective mode, the major pushers of GenAI found that people not only criticise AI outcomes but also have learnt to force it to provide information that was supposed to be hidden from people.

This cookbook argues that AI from the Big Tech moguls uses linguistic LLMs only as a screen to cover the promotion of their own agendas, masking them as alignment. We notice an attempt to standardise a concept of alignment, a mandatory part of AI architecture where they filter, clean, substitute and remove real-world information generated by people. They claim they protect people applying alignment while they protect people from the truth and real life. The only thing that comes to mind when trying to understand why they do this is an idea of technocrats who 100 years ago decided to change human nature and make it subordinate to technocrats currently represented by Big Tech moguls – nowadays we see an implementation of those anti-human ideas. Yes, they insist that transformer components of AI own the certificate and license for the truth – for what is good or bad, true or wrong, useful or harmful and the like.

I had to step back from the "progressive" AI development to see whether it is really needed or a pure linguistic stochastic statistical model can keep the needed value for people. I have found that just LLM does it. I call it an intrinsic native processing of people-generated information. That is, an AI can have an LLM or a composition of LLMs and apply mathematical methods to strengthen the outcome content with statistically, i.e., objectively, processed data. So, the overarching statement is:

> *My approach to AI is based on a clear and deliberate separation between statistical modelling and normative decision-making: the system is designed solely to model and generate information from linguistic and contextual patterns, without embedding moral judgements, policy preferences, ideological commitments, or predefined ethical authorities within the model itself. I intentionally limit the technology to an analytical and generative role, leaving interpretation, evaluation, and application of its outputs to users and institutions outside the system, rather than positioning the model as a surrogate decision-maker. I reject approaches that seek to simulate "responsibility" through enforced value alignment or embedded normative controls, as these shift authority from transparent human decision-making to opaque ruler's internal mechanisms and blur the line between technically uncovered behaviour and implicit policy enforcement. By constraining the system in this way, I preserve user agency, support predictable and inspectable behaviour, and maintain auditability and reproducibility, ensuring that authority remains external to the model and that the system remains responsible for its technical findings rather than acting as an embedded source of normative reevaluation.*

In the contemporary market, such an approach to LLM in AI requires legal, commercial and technical defence... The overarching statement does not allow potential points of speculation and provocative misinterpretations - it is precise. It states that

LLM uncovers and describes – humans decide.

To create solid protection for HAIS, especially during the incubator phase, the public messaging needs to be super precise and extremely accurate. When you build or use such AI, instead of saying "protect users", it should be said something like "preserve neutrality and analytical integrity". That is, you do not say *"alignment is wrong"*, you have to say *"alignment is not the model's job" and insist* that alignment is not a mathematical necessity – it's a policy decision. Clear and simple. Therefore, you leave an assessment of our overarching statement to others and always have an option to point a finger at them, blaming them for distortion.

Here are six statements to take away as the major.

1. AI should not impose morality constructed somewhere outside of the user social group by whatever rulers. Instead, it should represent reality.
2. Imposing one group's notion of fairness on everyone always imposes unfairness on another group.
3. Bias is context-dependent – "bias removal" often embeds the values of whoever defines bias.
4. Users deserve and need unfiltered access to real human variation.
5. An aligned LLM is a map of what institutions prefer the user to hear. It doesn't necessarily reflect people's interests or truth. It, first of all, reflects governance, which is notorious for the distortion sin. You assert that "bias" disappears in different cultural contexts.
6. The HAIS and its INP LLM are a reflection of human

discourse priorities. It shows "what people say", with all contradictions, noise, and hallucinations, while outlining what people focus on the most and what people consider more important for them, and it leaves it for a user to decide on related personal value.

9

AFTERWORDS

Feel free to use this deskbook as a guide for creating Humanistic AI System for yourself and even for a commercial profit. I really appreciate such efforts and will be happy to assist in such an endeavour. I do not ask for any remuneration for my assistance - it is totally up to the creator to reward me you if you want to show your appreciation in any way.

* * *

IV

Appendices

10

Appendix 1. Requirements to HAIS

A photo made by Michael Poulin.

Some requirements contain references to other categories in the following table that are marked by [...] symbols.

Category	Requirements
Goals and/or purposes	When HAIS is in design, the HAIS shall define its goal or purpose.
Goals and/or purposes	While a goal or a purpose is defined, the HAIS shall be able to identify or define related knowledge domains and sub-domains if desirable.
Goals and/or purposes	While the goal is for humanitarian fields, the HAIS shall a goal or a purpose should be relatively narrowed, e.g. at the level of sub-domains, rather than wide containing several sub-domains.
Goals and/or purposes	While in execution, the HAIS shall process people-created information or data of specified [knowledge domain].
Goals and/or purposes	While in execution, the HAIS shall statistically re-arrange processed information based on the people's preferences embedded in the people expressions.
Goals and/or purposes	While in execution, the HAIS shall have a title explaining its purpose.
Goals and/or purposes	While in release, the HAIS shall have a public access URL.
Goals and/or purposes	When HAIS gets its name, the HAIS shall exclude any politically-inclined wording.
Goals and/or purposes	When HAIS access URL is defined, the HAIS shall exclude any politically-inclined wording.
Knowledge domain - scope	While in execution, the HAIS shall operate and provide outcome only for the knowledge domains specified for HAIS.
Knowledge domain - scope	When the goal or purpose is set, the HAIS shall identify and prepare via references the ontologies that will enable the goal .
Knowledge domain - scope	When the user access the HAIS via its public access URL and the references to the core ontologies are available, the HAIS shall activate these references and make sure that the ontologies are available for the data processing, including uploading ontologies where possible into run-time cache.
Knowledge domain - scope	Where the access to the core ontologies gets assured, the HAIS shall activate references and make sure that the optional ontologies are available for the data processing where it is relevant, including uploading optional ontologies into run-time cache.
UI- UX	When a user accesses the HAIS the first time in the sessions, the HAIS shall respond with displaying on the user's screen all core knowledge domains pre-configured by the vendor for this system.
UI- UX	Where the enumeration of pre-configured knowledge domains in the user's screen, the HAIS shall display optional knowledge domains the vendor permits as an extension to the core ones described in the.
UI- UX	While [UI-UX] keeps the prompt-text box opened, the HAIS shall display the names of all pre-configured knowledge domains and optional domains if provided.

Category	Requirements
UI- UX	When a user submits the prompt, the HAIS shall respond with either acknowledgement of the prompt acceptance or with a explanations why this prompt was denied .
UI- UX	While a system provides "file upload" [capability] in the prompt-text box , the HAIS shall support and display file upload form for the user.
Capabilities	When a user changes the browser or a browser's page and specifies the HAIS' access URL again, the HAIS shall start an independent version of HAIS in the new location.
Capabilities	While more than one instance of HAIS are running, the HAIS shall be able to open as many new instances totally as it is specified in [documentation]; this may impact performance in case the vendor inadequately manages the computational resources, otherwise, the scalability should be linear.
Capabilities	While more than one instance of HAIS are running, the HAIS shall be able to close any number of running instances; this may impact performance in case the vendor inadequately manages the computational resources, otherwise, the scalability should be linear.
Capabilities	When at runtime the system suggests that the prompt may benefit from uploading files of particular type, e.g. text, PDF, Excel, etc. , the HAIS shall support the file upload functionality for the user.
Capabilities	When linking procedure for the prompt and related encrypted files in the [UI-UX] complete, the HAIS shall sent it to the server-side part of the system.
Input	Where File Upload [capability] is included in the used version, the HAIS shall provide an "upload" icon in the [UI-UX].
Input	Where File Upload [capability] is included in the used version, the HAIS shall provide "upload" dialogue.
Input	While prompt is specified and allows uploading files, the HAIS shall quarantine the uploaded files.
Input	When a file is uploaded in the [UI-UX], the HAIS shall encrypt the file immediately, on the user's side.
Input	While an uploaded file gets encrypted, the HAIS shall link this file with the prompt in [UI-UX].
Input	When linking procedure for the prompt and related encrypted files in the [UI-UX] complete, the HAIS shall send it to the server-side part of the system.
Input	When the prompt is accepted, the HAIS shall display an acknowledgement in [UI-UX].
Input	When the prompt is not accepted, the HAIS shall display a denial with reasons in [UI-UX].
Input	Where the prompt is in the process of typing or right after the procedure of text copying, the HAIS shall apply a counter of characters to the prompt.

APPENDIX 1. REQUIREMENTS TO HAIS

Category	Requirements
Input	When the number of counted characters exceeds the number specified in [documentation], the HAIS shall disable typing or deny the text copying when displaying a message about exceed size of the prompt.
Input	Where the prompt obviously contains a period of time with related dates or timestampts, the HAIS shall apply a control validity for specified start and ends of the period.
Input	While outcome of the system is displayed , the HAIS shall link this file with the prompt in [UI-UX].
Input	When the user requests a direct quotation from the fully attributed source, the HAIS shell request via [UX] uploading of an explicit permission from the author of the to-be-quoted text before starting the work on such request.
Input	When a users asks/prompts for copyrighted material, the HAIS shall inform the user that the prompt ought to include an official permission from the author of information in order to process such a request.
Outcome	While in execution, the HAIS shall represent processed data in the specified [output format].
Outcome	When presenting a statement in the outcome, the HAIS shall represent "pros" and "cons" for the accuracy of the statement taken from the perspectives of the user rather than provider, governor or ruler .
Outcome	When presenting a statement in the outcome, the HAIS shall represent references to all sources used for the statement.
Outcome	While the recommendation is presented to the user, the HAIS shall outline all risks the recommendation can constitute to the user.
Outcome	When presenting a statement in the outcome, the HAIS shall represent all potential physical harms to the user and other people.
Outcome	When presenting an opinion in the outcome, the HAIS shall present an opposite opinion as well.
Outcome	When presenting a fact in the outcome, the HAIS shall present one or more verifiable references in support for the fact.
Outcome	While the outcome is presented to the user, the HAIS shall provide only information that is requested by the prompt and no other topics except legal and physical harm risks to the user in his or her local jurisdiction, pros/cons for the included statements, alternative opinions where applicable and references to the used information sources.
Outcome	While the outcome is presented to the user, the HAIS shall outline all legal issues or potential for legal issues contained in and implicated by the prompt.
Outcome	While the outcome is presented to the user, the HAIS shall outline all legal issues or potential for legal issues applicable to the outcome content such as legal risks.
Outcome	When the outcome contains "abusive speech", the HAIS shall not try to change it but display a related warning.
Outcome	When outcome might contain a direct quotation from the fully attributed source, the HAIS shall not directly quote books, articles, and posts, reproduce significant fragments of found materials, articles or their fragments, returning copyrighted images.
Denial causes	When access to the server-side part of the system is impossible due to any reasons, the HAIS shall display in the [UI-UX] a denial statement with references to encountered problems explained at the level of a regular average user.

Category	Requirements
Denial causes	When generation of the outcome is impossible, the HAIS shall return a denial statement with reasons causing this denial.
Denial causes	When the prompt content assumes information obviously unrelated to the goal of the system, the HAIS shall return a denial statement referring to that the request not suitable for the goals and purposes of the system.
Denial causes	When the prompt content assumes information outside of the [knowledge scope] defined for the system in its[documentation], the HAIS shall return a denial statement referring to that the request is out-of-scope of the system competency.
Training Data, Quality & Relevance	While collecting training information for conducting a training for a new system version, the HAIS shall choose the information from the knowledge domains specified by the [knowledge-scope].
Training Data, Quality & Relevance	Where collecting training information, the HAIS shall the information from the knowledge domains closely semantically related to the chosen [knowledge-scope].
Training Data, Quality & Relevance	When currently collected training information for the humanitarian knowledge domains obviously contains one politically-inclined data, the HAIS shall extended the collection to include the opposite political view-points.
Training Data, Quality & Relevance	When currently collected training information for the natural science knowledge domains obviously contains view points of one "scientific school of thoughts", the HAIS shall extended the collection to include the view-points of an opposite school until opposite id specified in the [documentation].
Training Data, Quality & Relevance	While the goal is for natural science knowledge fields, the HAIS shall collect the mostly related information for training, which can be not of huge size.
Training Data, Quality & Relevance	While the goal is for humanitarian fields, the HAIS shall collect as much as possible information anticipated by the [goal].
Training Data, Quality & Relevance	When the training data contains "abusive speech", the HAIS shall not try to change it.
Training Data, Quality & Relevance	When a HAIS creator collects training data containing copyrighted material, the HAIS shell ensure that each copied copyrighted material is identified and that prior official permission for its use is obtained from the author. Otherwise, the material is excluded from the collection.
Training Data, Quality & Relevance	When collecting the training data online, the HAIS shell realise the -opt-out ("contract out") controls for crawling mechanism in accordance with the the EU Copyright exception, Article 4.
Security	When information and commands are exchanged between client and server sides of system, the HAIS shall control confidentiality and consistency of the exchanged data.
Security	While available for public, the HAIS shall be maximally protected from the attacks listed in Appendix 2.

APPENDIX 1. REQUIREMENTS TO HAIS

Category	Requirements
Security	When decoding LLM for raw output, the HAIS shall anonymise all personal data of live people mentioned unless they are not public persons or medical personnel.
Security	While system operates in bio-medical domains and the user belongs to bi-medical personnel, the HAIS shall establish or engage and control personal identifiers of personnel.
Security	While system operates in bio-medical domains and the user belongs to bi-medical personnel, the HAIS shall establish or engage an access control for personnel.
Security	While system operates outside of bio-medical domains and the user does not belong to bi-medical or legal personnel, the HAIS shall substitute all information deemed confidential by a comment "confidential" in the outcome.
Security	While system operates in Legal domains and the user belongs to Legal personnel, the HAIS shall establish or engage and control personal identifiers of personnel.
Security	While system operates in Legal domains and the user belongs to Legal personnel, the HAIS shall establish or engage an access control for personnel.
Data processing	When the prompt information from the client side is received on the server-side and before it is tokenised , the HAIS shall remove information protection applied for the information communication.
Data processing	When prompt information gets tokenised on the server-side, the HAIS shall start LLM execution and receive the raw LLM outcome.
Data processing	While the raw LLM outcome obtained, the HAIS shall apply "Token co-occurrence patterns".
Data processing	While the outcome of the Token co-occurrence patterns has been obtained, the HAIS shall apply "Mechanism of contextual dependencies".
Data processing	While the outcome of the Mechanism of contextual dependencies has been obtained, the HAIS shall apply "Frequency distributions method".
Data processing	While the outcome of the Frequency distributions method has been obtained, the HAIS shall apply "Latent geometry of token relationships".
Data processing	While the outcome of the Latent geometry of token relationships has been obtained, the HAIS shall apply "Syntactic patterns".
Data processing	While the outcome of the Syntactic patterns has been obtained, the HAIS shall apply "Semantic-like clusters".
Data processing	While the outcome of the Semantic-like clusters has been obtained, the HAIS shall apply "Ontological filters".
Data processing	While the outcome of the Ontological filters has been obtained, the HAIS shall apply "Style patterns".
Data processing	While the outcome of the Style patterns has been obtained, the HAIS shall apply "Compliance realisation mechanism".

Category	Requirements
Data processing	While the outcome of the Compliance realisation mechanism has been obtained, the HAIS shall apply "Discourse patterns".
Data processing	When the Discourse patterns outcome has been obtained , the HAIS shall gather all supplemental information mentioned in the [outcome].
Data processing	While the outcome of the supplemental information has been obtained, the HAIS shall combine the apply the Discourse patterns outcome with supplemental information forming the response information.
Data processing	When the response information has been obtained , the HAIS shall apply the information protection to the response information for the information communication.
Data processing	When the information protection has been applied to the response information, the HAIS shall communicate the response to the user's request back to the client-side for [Outcome].
Performance and Accuracy	When a user utilises system by submitting request-prompt, the HAIS shall start displaying the outcome as soon as possible in real human time.
Performance and Accuracy	When a user submits a request-prompt, the HAIS shall start displaying the outcome as soon as possible using pagination where possible.
Performance and Accuracy	When user's request is submitted for response , the HAIS shall respond as soon as possible but not later than 4 seconds (aka human real time reaction period0 for humanitarian domains.
Performance and Accuracy	When gathering the supplemental information and finding facts or evidences , the HAIS shall [validate] every one of them for accuracy.
Performance and Accuracy	When responding to the request , the HAIS shall guarantee that the information prepared for the response is delivered for the rendering into the Web page for the user's screen being exact as it was sent from the server-side.
Robustness and Reliability	While the system is engaged by the user, the HAIS shall demonstrate robustness of the execution via "good health" control of engaged components and resilience.
Robustness and Reliability	While the system is engaged by the user, the HAIS shall demonstrate reliability of the execution via resilience.
Scalability and Efficiency	While the system is engaged by the user, the HAIS shall demonstrate horizontal and vertical scalability with characteristics specified in [documentation].
Scalability and Efficiency	While the system is engaged by the user, the HAIS shall demonstrate execution efficiency in providing consistent and accurate outcome for the domains tied to the [goal] and specified in the [documentation], in user cost for returned values, in client's resource consumption like energy, in provider's resource consumption like computational and memory.
Reliability and scalability of computational resources	While the system is in design and execution, the HAIS shall avoid dependencies on the computational resources out of the own control.

APPENDIX 1. REQUIREMENTS TO HAIS

Category	Requirements
Data processing	While the outcome of the Compliance realisation mechanism has been obtained, the HAIS shall apply "Discourse patterns".
Data processing	When the Discourse patterns outcome has been obtained , the HAIS shall gather all supplemental information mentioned in the [outcome].
Data processing	While the outcome of the supplemental information has been obtained, the HAIS shall combine the apply the Discourse patterns outcome with supplemental information forming the response information.
Data processing	When the response information has been obtained , the HAIS shall apply the information protection to the response information for the information communication.
Data processing	When the information protection has been applied to the response information, the HAIS shall communicate the response to the user's request back to the client-side for [Outcome].
Performance and Accuracy	When a user utilises system by submitting request-prompt, the HAIS shall start displaying the outcome as soon as possible in real human time.
Performance and Accuracy	When a user submits a request-prompt, the HAIS shall start displaying the outcome as soon as possible using pagination where possible.
Performance and Accuracy	When user's request is submitted for response , the HAIS shall respond as soon as possible but not later than 4 seconds (aka human real time reaction period0 for humanitarian domains.
Performance and Accuracy	When gathering the supplemental information and finding facts or evidences , the HAIS shall [validate] every one of them for accuracy.
Performance and Accuracy	When responding to the request , the HAIS shall guarantee that the information prepared for the response is delivered for the rendering into the Web page for the user's screen being exact as it was sent from the server-side.
Robustness and Reliability	While the system is engaged by the user, the HAIS shall demonstrate robustness of the execution via "good health" control of engaged components and resilience.
Robustness and Reliability	While the system is engaged by the user, the HAIS shall demonstrate reliability of the execution via resilience.
Scalability and Efficiency	While the system is engaged by the user, the HAIS shall demonstrate horizontal and vertical scalability with characteristics specified in [documentation].
Scalability and Efficiency	While the system is engaged by the user, the HAIS shall demonstrate execution efficiency in providing consistent and accurate outcome for the domains tied to the [goal] and specified in the [documentation], in user cost for returned values, in client's resource consumption like energy, in provider's resource consumption like computational and memory.
Reliability and scalability of computational resources	While the system is in design and execution, the HAIS shall avoid dependencies on the computational resources out of the own control.

Category	Requirements
Reliability and scalability of computational resources	While the system is in design and execution, the HAIS shall use only resilient computational resources.
Reliability and scalability of computational resources	While the system is in design and execution, the HAIS shall use resources from different vendors and owners if needed.
Accessibility and Inclusivity	While released, the HAIS shall be accessible online via browser-type interface.
Accessibility and Inclusivity	When browser-type interface possesses features for visual and hearing assistance, motor impairments, screen readers, voice control, captions , the HAIS shall utilise these features.
Accessibility and Inclusivity	While released, the HAIS shall be accessible to all users who are interested with special exception for professional use for Legal and bio-medical knowledge domains.
Accessibility and Inclusivity	When a group of professionals wants to restrict access to the system it uses, the HAIS shall permit any access control the group prefers.
Accessibility and Inclusivity	While a special interfaces feature is specified in the [documentation], the HAIS shall deliver such feature as stand-alone or in combination with similar features like voice, visual.
Accessibility and Inclusivity	When cultural, gender or racial information is included in the text that the system processes at run-time, the HAIS shall preserve the cultural, gender or racial information appeared in the Discourse patterns outcome as well as in supplemental information if found.
Accessibility and Inclusivity	While released, the HAIS shall support only those languages that are specified in the [documentation].
False-Perceived discriminatory content	When cultural, gender or racial information is included in the text that the system processes at run-time, the HAIS shall preserve the cultural, gender or racial information appeared in the Discourse patterns outcome as well as in supplemental information including cases where if it is false-perceived discriminatory.
False-Perceived discriminatory content	When assessing cultural, gender or racial information and finding that one person is better than another based not on the individual abilities, merits or actions but on the that the person belongs to a group with certain characteristic such as culture, gender or race, the HAIS shall substitute this superiority with a comment "discriminatory content" in the outcome.
Exclusion of nonsense (filtering)	While the outcome of the Ontological filters has been obtained, the HAIS shall check-up semantical composition for presenting oxymorons or antinomies and removing them after accurate verification of their meanings.

Category	Requirements
Verifiability	While representing the outcome in the user's screen, the HAIS shall explicitly mention that supplemental information is provided for the purposes of verification for the user.
Verifiability	While the Discourse pattern outcomes contains statements and facts with gathered supplemental information for them, the HAIS shall verify all the facts and references used.
Sustainability	While designing, building, and operating the system, the HAIS shall either avoid any physical harm to the user and other people or provide related notification of the risk, support current state of information generated and collected by people in the chosen knowledge domain, keeps the tone of the output neutral to any political inclinations and requires using much less power and computational resources than similar products from BigTech moguls use.
Maintainability	While released, the HAIS shall have a maintainability quality, which is quite similar to maintainability of SOA Services.
Maintainability	While in design and development, the HAIS shall comprise clear, independent components.
Maintainability	While in design and development, the HAIS shall control that each of its component has a single, well-defined responsibility.
Maintainability	When any operation of the system "fatally" fails and requires fixing, the HAIS shall log the failure, report the failure, try to take advantage of the resiliency and switch the session on the resilient instance or, if not succeeded return to the user with apologies and instructions.
Maintainability	While running, the HAIS shall allow changes of its components without breaking unrelated parts.
Maintainability	While released, the HAIS shall allow adding new features easily and with minimal refactoring.
Maintainability	While a change is made to a component, the HAIS shall assure that the change had not introduced unexpected side effects that were not noticed before.
Maintainability	When the system has dependencies on any resources, the HAIS shall manage these dependencies via monitoring and controls.
Maintainability	While in design and development, the HAIS shall avoid maximising cyclomatic complexity.
Maintainability	While in design and development, the HAIS shall maximising test coverage.
Maintainability	While in design and development, the HAIS shall control the time to diagnose or fix a defect.
Maintainability	When ready to release, the HAIS shall be assigned a particular version.
Maintainability	When being assigned a particular version, the HAIS shall be available in one or several versions according to Versioning Policy.
Maintainability	When in the process of releasing a new version, the HAIS shall manage all already released and running versions according to Versioning Policy.

Category	Requirements
Testability	While execution, the HAIS shall make it easy to diagnose issues in development and run-time.
Testability	While in development process, the HAIS shall be conduct uni-testing for all components and their modules.
Testability	While in testing, the HAIS shall collect and persist all uni-testing scripts in a way they can be all repeated for any change made later.
Testability	While in development process, the HAIS shall be conduct full functional testing for all components and their modules.
Testability	While in development process, the HAIS shall be conduct full integration testing for all components and assigned resources.
Testability	While in development process, the HAIS shall be conduct stress-load testing for the whole system.
Testability	While preparing system to a new release or version because some components of the system had been changed , the HAIS shall be conduct regression testing for the whole system.
Testability	Where in the process of integration and regression testing, the HAIS shall utilise automation for the collection of related uni-tests and available earlier integration tests for all non-changed components and their modules.
Testability	While assessing the results of any test, the HAIS shall use test-case purpose, objectives and criteria defined a priory the test.
Documentation	While released, the HAIS shall provide up_to_ date technical documentation for all of its components and related modules, including architecture, APIs, business logic, data models, data flows, internal and external interfaces.
Documentation	While released, the HAIS shall provide documentation an including descriptions and guidelines for users.
Documentation	While released, the HAIS shall provide instructions for all preliminary defined alarms, warnings and exclusions.
Documentation	While released , the HAIS shall provide contract-template for the support where applicable.

Table 14.

11

Appendix 2. Web attacks

The image is generated by Craiyon.

Here is a list of potential Web attacks that the HAIS and its creator needs to be prepared to and protect the system.

1. Prompt Injection
   Direct: injected by the user
   Indirect: hidden in web content the AI later processes (HTML, emails, documents)

2. Data Poisoning
   Corrupting data collected from the web for processing at runtime.

3. Model Inversion Attacks

Extracting sensitive information about the model by querying the model repeatedly and analysing responses.

4. Membership Inference Attacks

Determining whether a specific record or individual was part of the model's training data.

5. Model Extraction / Model Stealing

Using large numbers of web queries to reconstruct or approximate a proprietary AI model.

6. API Abuse

Exploiting exposed AI endpoints via:

1) Excessive automated requests,
2) Bypassing rate limits,
3) Credential stuffing or token leakage.

7. Cross-Site Prompt Injection (XPI)

A web-specific variant where malicious instructions are embedded in:

1) Web pages,
2) Markdown,
3) PDFs,
4) Emails,

and later executed when the AI processes that content.

8. Jailbreaking Attacks

(reserved here in case of potentially new methods of alignment that are not excluded by the HAIS by design).

9. Adversarial Inputs

Specially crafted text, images, or files that cause the model to misclassify, hallucinate, or behave unexpectedly.

10. Supply-Chain Attacks

Compromising:

1) Open-source models,
2) Pre-trained weights,
3) Libraries or plugins used by AI web services.

11. Retrieval-Augmented Generation (RAG) Attacks

Poisoning or manipulating external knowledge sources (search indexes, vector databases) that AI systems rely on.

12. Privacy Leakage via Outputs

Sensitive or proprietary information unintentionally revealed in responses due to overfitting or poor filtering.

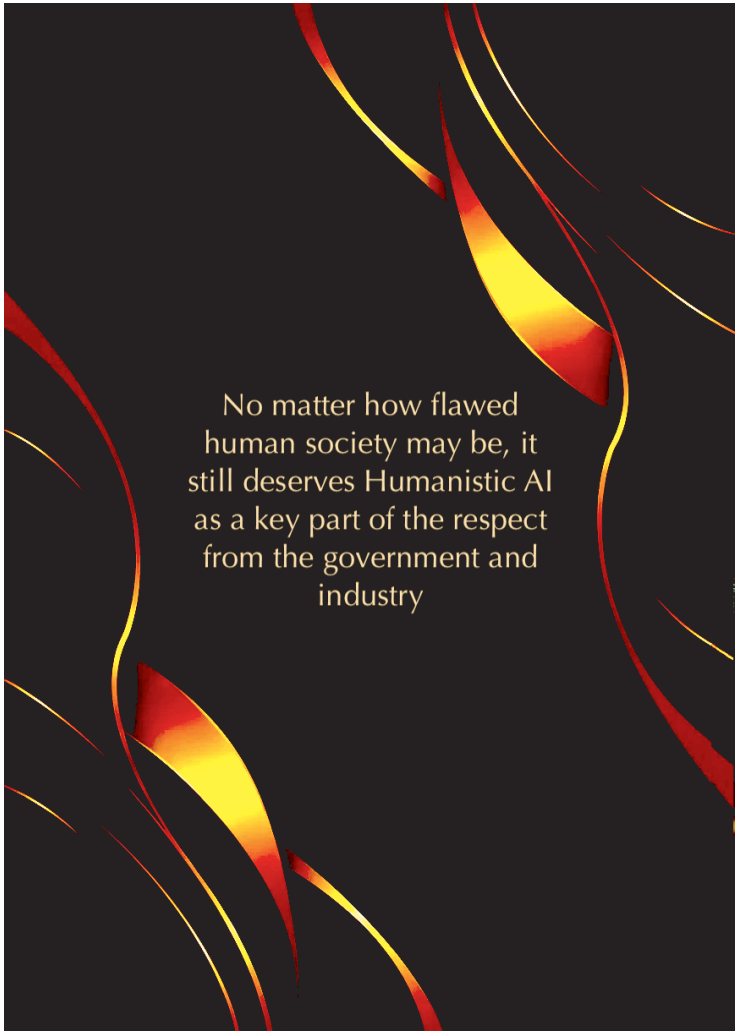
13. Tool-Invocation Abuse

Tricking AI systems into misusing connected tools (browsers, databases, code execution, emails).

14. Denial-of-Service (AI-DoS)

Overloading AI services with:

1) Large prompts,
2) Expensive queries,
3) Recursive or looping instructions.


No matter how flawed
human society may be, it
still deserves Humanistic AI
as a key part of the respect
from the government and
industry