Ethics in the future of AI - a philosophical perspective for collaborative human and machine learning know-what and know-how

Iustina Neagu
Institute of Philosophy and Psychology "Constantin Rădulescu Motru"
Romanian Academy
Bucharest, Romania
iustina.neagu@yahoo.com

Ciprian Daniel Neagu
School of Computer Science, AI & Electronics
University of Bradford
Bradford, UK
d.neagu@bradford.ac.uk

Marius Augustin Drăghici
Institute of Philosophy and Psychology "Constantin Rădulescu Motru"
Romanian Academy
Bucharest, Romania
mariusdraghiciinstitut@yahoo.com

2025 1st International Conference on Computational Intelligence Approaches and Applications (ICCIAA) | 979-8-3315-2365-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICCIAA65327.2025.11013343

***Abstract*— In this paper we argue with pragmatic review of scientific evidence why current approaches to define and evaluate the debatable Artificial General Intelligence claims need fundamental philosophical background, knowledge and skills for specialists in the era of Generative AI to address ethical and logical challenges, particularly with the view of risks and costs associated with their use and developments. The article proposes the Kantian perspective of ethics and its potential in the new AI developments towards Artificial General Intelligence as trustful part of a formal meta-design framework, and fundamentally humanistic ways of implementation and evaluation, with examples of Generative AI impact in academic research misconduct creation, evaluation and dangers, such as plagiarism and decision errors.**

Keywords*— *Artificial General Intelligence, Kantian Ethics, Machine Learning, Meta-design, Ethical AI, Entropy

I. INTRODUCTION

Generative Artificial Intelligence (GenAI) developments, successes and challenges are opening, with continuous business pressure, both opportunities and risks for the human society, with practical manifestations in our lives (e.g. in social media, education, industry and economy, healthcare, job market and regulations among main domains). Moreover, the discrepancies and criticisms for GenAI technologies creation and adoption are now a fashionable presence in core and multidisciplinary scientific research due to increasing speeds of impactful progress and applications, mainly because of abilities and large spectrum developments in simultaneous industrial revolutions [1]. However, this ongoing and growing pressure of AI occurrence in everybody's existence comes with signals of incomplete, inconsistent or impractical coverage of key concepts and processes that link AI and humanity through technology, with the visible challenge of epistemology and ethics in decision making. Powerful examples come from industries and experts claiming positive avenues for humanity that reflect fast and successful GenAI adoption in creative industries, assistance, healthcare, while warnings, alarms and efforts to measure risks (to humanity and environment) are also raised by experts and regulators.

The stakeholders in AI developments and technological advances that make nowadays the breaking news are all: the public, the industry, the users and beneficiaries of digital technologies, the scientists and their audience. However, the rapid developments of AI technologies are currently driven by consumerist and industrial rationale, while research and regulatory policies study risk assessment - where "big data" were the fuel of the technology advances just 20 years back, now are GenAI models: nothing more than "big data-fueled engines". Such progress from data and information to knowledge comes at high costs of energy and trust.

Moreover, we are technically still progressing within Data-Information-Knowledge-Wisdom (DIKW) pyramid of resources [2][3][4], potentially giving voice to Data and Information through multimodal "Chatbot" Large Language Models (LLMs) built on large though inconsistent Knowledge representations. We eyewitness and participate in yet another industrial revolution (IR) that captures footprints of all resources [1]: mechanisation – IR1.0, energy (electrification - IR2.0), large scale automation and production (industrial robots, cyber-physical systems – IR3.0), digitalisation (data-centric – IR4.0) personalisation (human-centric – IR5.0), and humanisation – IR6.0, building on previous successes, but aiming to progress not just in the material world, now more within human perception coordinates.

Therefore, our scientific challenge and proposed approach refer to: *how can one manage and get everybody and everything around a mindful dialogue table with robust foundations and trustful outcomes*? Interestingly, this challenge has been seen in the past, and solutions require pragmatic multidisciplinary research to clarify the involved concepts, entities, process updates, solutions and vision for progress: the philosophical perspective. The holistic power of ethics, logic and knowledge in human reasoning is necessary at these stages of development of GenAI with presumed human features to identify commonalities and differences for the actors in the era of GenAI towards Artificial General Intelligence (AGI). The paper proposes consequently a pragmatic review of state of the art of GenAI progresses, strengths and weaknesses, as a foundation for proposing an update of AGI design by the inclusion of Kantian ethics. The theoretical framework will be argued within the Data-Information-Knowledge-Wisdom (DIKW) hierarchy with the proposal of the ethical meta-design of GenAI. Examples to support the proposed framework refer to case studies on positive and negative impacts of GenAI in research paper writing from the consistent perspective of benefits and risks associated with paper mills, plagiarism detection, false claims.

The remaining content of the paper consists in: Section 2 presents the background, motivation and methodology. Section 3 is a pragmatic exploration of concepts and proposes a philosophical perspective of co-existence of artificial intelligence and human wisdom, their similarities, differences and testing through DIKW lens. Section 4 argues the potential advantages brought by the Kantian vision of pure reasoning and ethics for the human and AI agencies, and proposes a holistic, integrative framework for Ethical GenAI meta-design. Case studies on GenAI use and impact to scientific papers are argued with references, arguments, examples, and

C.D. Neagu acknowledges partial funding from FS-2324-20-26 CoMAP funded by the Royal Academy of Engineering under the Frontiers seed funding - FY2324 - Tranche 2.

work in progress in Section 5. The paper ends with a summary of reported work and future research directions in Section 6.

II. A Pragmatic Approach of the State of the Art

A. Background and Motivation

AI generalised adoption arrived in the scientific "ivory tower" with opportunities (innovation, growing access to larger resources, limitless applications with impact) and challenges (academic misconduct, deep fakes, hallucinations and lack of trust). The domain of scientific publishing is nevertheless affected heavily [5]. Nowadays more and more papers are retracted from journals on a number of issues. However, this has potential impacts in "correct" papers that included retracted articles in their scientific argument. This snowball effect is in itself an argument of risks commonly associated with AI misuse. Even more confusing are "fallacies"-generating corners of the confusion matrix in decision making created by false positives and false negatives of the potential academic misconduct, misinterpretation, lack of transparency or process errors. Scientific publishing, including innovation, communication and checks, is a domain of continuous pressure for decision makers and stakeholders (co-authoring team members, publishers, reviewers, editors, academic institutions, international ranking systems).

This argument has for this paper a double rationale: to support the pragmatic review the authors provide by selecting references of meaningful content and source with direct evidence of confident content, and to provide the body of work for illustrative examples as case studies. Consequently, this paper, while avoiding the "systematic review" (motivated partly by the presence of potentially hallucinated, and difficult to trust and follow scientific outputs "within the era of retracted and GenAI-generated resources" [6] that motivates the case studies), proposes to use a rather pragmatic identification of ethically relevant, efficient, essentially deontological resources and references for the DIKW evaluation using philosophically motivated perspectives.

Our pragmatic review builds on the existing dedicated DIKW concepts and developments connected in the industrial revolutions. As in any industrial revolution, what Science has proposed, and Engineering implements, becomes an object of Management with its implications (Performance, Costs, Risks, Governance). This shows clearly the current stage with reference to the DIKW pyramid, due to current progresses in AI: the domains of 1) Data {Science, Engineering, Management} that are long term established; 2) Information {Science, Engineering, Management} that are long term established. 3) The Knowledge domains are though more debatable at least semantically: Knowledge Science comes with traditional perspectives of epistemology and modern views of knowledge science [7]. However, about intelligence and wisdom, the domains are yet to be claimed! Consequently, where are the foundations of Intelligence Science, Intelligence Engineering and Intelligence Management domains? With this motivation and challenge, the paper aims, reflecting on the parallel developments in DIKW hierarchies particularly within the framework of modern and contemporary industrial revolutions, to address the open field of Intelligence/Wisdom Science/Engineering/Management debate, and proposes, within the overall perspective of Artificial General Intelligence, a contemporary, critical and fresh contribution of philosophy: Kant ethics and AI [8][9].

B. Methodology

This paper argues with references to trustful logical and statistical evidence why current approaches to define and evaluate debateable AGI need fundamental philosophical background and knowledge for specialists in the era of Generative AI to address ethical and logical challenges particularly with the view of risks associated with their use and developments towards Artificial General Intelligence. We identify through critical analysis the key challenges and open fields for scientific progress in the DIKW evolution as part of industrial revolution series. We bring the Kantian perspective of ethics and its potential in the developments towards Artificial General Intelligence, and propose the collaborative integration of humanistic ways in the design, implementation and evaluation of AGI samples. For evidence and with relevance to the scientific debate, this paper is building the innovative perspective on examples from GenAI impact in academic research papers: misconduct creation, propagation, evaluation, impact and dangers, as well as fallacies and errors.

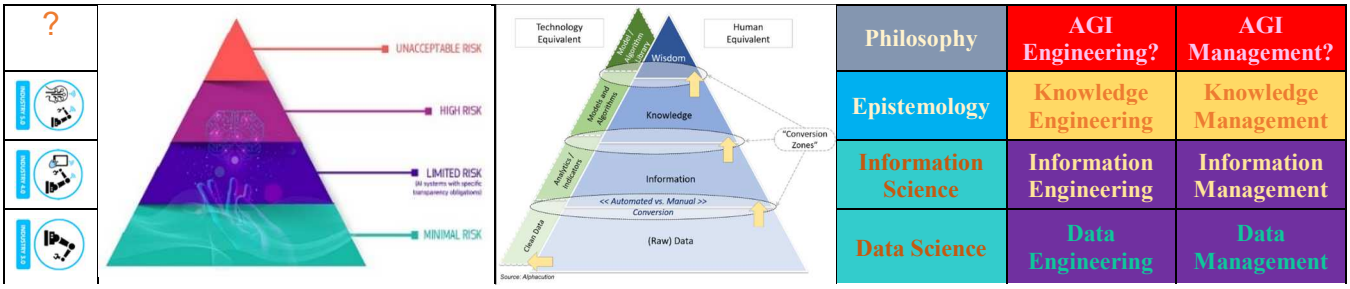

Fig. 1. DIKW, Automation, and AI Risks Hierarchies with Science, Engineering and Management and Industrial Revolutions Perspectives (images [10][11])

As a first step, building from the potential impact of academic misconduct in research paper production and publication, the authors are focused on a pragmatic review approach in the paper. The approach in avoiding systematic review is potentially influenced by current and past under development retractions and misconduct cases that are growing uncontrollably. By using a pragmatic identification of efficient, essential references for DIKW-based evaluation of AI vs human intelligence in current times, this paper focuses on the „pragmatic" mainstream of academic publications [5][6]. As the second step, a critical analysis of the consistency of coverage of DIKW concepts within an industrial revolution era (of GenAI towards AGI). The vision of Human Wisdom (HW) vs Artificial Intelligence (AI or soft/ware wisdom: SW, as a consistent mirror of AGI within DIKW) supports the proposal for the Intelligence Science, Engineering and Management framework attempt produced in relation to DIKW pyramid (Fig 1): firstly by addressing the argument of Intelligence vs Wisdom Science with regard to concepts; then Intelligence Engineering with process meta-design arguing the necessity of the innovative presence of Kantian ethics with *a priori* and *a posteriori* presence; and

finally Intelligence Management/Governance with identification of metrics for the similarities and differences of concepts. The pragmatic review will produce the arguments in relation to DIKW concepts, their definitions of potential measurements from data-centric vs human-centric views.

In the third step the proposal for AGI meta-design with Kantian ethical perspectives for holistic *a priori* and *a posteriori* contributions are described and supported with examples from research publications troubles in GenAI era.

III. A Pragmatic Review of Relevant Concepts through Challenges

A. GenAI testing from DIKW hierarchy perspectives

AI key concepts and definitions used today are not consistent, nor acknowledged consistently although data, information, knowledge, intelligence, agents/agentic machine and deep learning, reasoning, ethics, risks, bias are nowadays used by all. In fact, such concepts, individually and also together (as in the DIKW hierarchy), have different meanings for various stakeholders, and even scientists alone (Fig 1), and of course general public (businesses, users, legislators). Moreover, this challenge produces additional waves of problems, confusions, consequences and impact from, to and between the scientific, engineering, management, legal and users' audiences. Consequently, AI advances and counter-actions from society become and create themselves additional points of interest, debates, crises, technological or legal cases.

The impact of AI and related representations (ontologies), reasoning (epistemologies) and risks (including ethical aspects) are already present and their legal aspects captured in new policies and legal acts (e.g. EC AI Act [10]). Where science and legal approaches become part of the technological landscape, evaluation requires identification of features and risks – for (artificial) intelligence systems – and this nowadays could be done by human perception or legal evaluation.

The surprising facts are that in such management and legal attempts, the concepts are not defined nor clarified for evaluation purposes at least. The progressive motion from 2006 "data is the new oil" metaphor of British mathematician Clive Humby (as oil powers the modern industrial revolutions, data fuels digital technologies as economy engines to generate value) comes at additional costs of energy, risks and uncertainties. However, what in 2006 was data (as computational values), just years later was information – representations of facts and objects for human perceptions (images, sounds, videos, emotions, tastes, etc.) becoming processable by specialised hardware and software. In other words, data is for computing machines what information is for humans (initially though, as now both, data and information are processable by all parties, with plenty examples the machine is faster than humans in processing larger amounts of (big) data + information. Consequently, the first risk of AI overtaking humans is confirmed by narrow applications: chess, medical scan processing, translations, searches, summarisations, analytics and visualisation.

With Large Language Models and their abilities to capture and process tokens, words, and patterns, we are nowadays the eyewitnesses, contributors and beneficiaries of Generative AI systems becoming the new oil for the human society. The rapid progress from data, then through information to knowledge "fuels" comes at the costs of: energy consumption; uncontrollable "fuel" quality (featuring provenance, algorithms, prompt engineering issues); non-deterministic outputs given the continuous, dynamic change of the knowledge-base used to produce and deploy Gen AI, including hallucinations, sourcing intellectual rights and incontrollable risks. Furthermore, GenAI has captured the attention of users by reacting to human perceptions, by text, image, sound, video, and coming more sensorial outputs. GenAI demonstrates additional risks by mimicking reactions that, if produced by humans, may be classified as intelligent! GenAI started passing various intelligence tests: Turing Test, CAPTCHA, IQ, academic and professional examinations.

A natural development, the original Turing Test (TT) [12] and their variants [13] are acknowledged by addressing AI manifestations towards what humans perceive as intelligent behaviour and therefore treating AI from the human perception frontend, technically allowing AI perception with humanist/humanisation attitude. As AI evolves in terms of resources and functions, the measure of (artificial) intelligence is therefore a must in the general scientific journey towards acknowledgment of AI detection [14][15]. CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) [16] is a measure used to distinguish human users from automated bots for cyber security aims. While TTs use human judge, CAPTCHA decision is machine-controlled. IQ tests, introduced to evaluate human cognition and reasoning initially, have shown advances acknowledged in the way of measuring AI IQ [14], either by software or by social tests and exams passed by GenAI [17][18][19].

What is the main issue though? Actually, TT and IQ tests see AI from a human perspective, while CAPTCHA tries to differentiate them by simple tasks – and all tests are continuously challenged and updated though. But is this approach correct to measure AI in the way human intelligence is measured? The EU AI Act already accepts and acknowledges the fact that AI developments cannot be addressed at source and therefore the entire dynamically and fast progressing AI technologies presence in human society (GenAI, AGI/ASI, Singularity, master-slave dilemma) cannot be addressed but just by their risk of adoption and presence.

B. Ethical AI within Kantian Ethics perspectives

The paper captures and builds on the current developments proposing, through the Kantian deontological perspective applied to the identification of the key DIKW concepts (for the science streams), a fundamental view and potential metrics (for the engineering and evaluation management streams). This perspective, that promises a theoretical holistic and trustful approach to solution developments for ethical AI, starts taking shape building, among other things, on recent works [9][20][21][22] in addition to the fundamental Kantian works: Critique of Pure Reason, Critique of Practical Reason, Prolegomena, Metaphysics of Morals, mentioned in the next sections below. The paper also raises the issue of common and different features, properties and functionalities between (Human) Wisdom vs (Artificial General) Intelligence (AGI) such as the debatable utilitarian, biased, limited and obsolete "data-centric" approach with "chat bot" behaviours passing Turing Tests without passing the AGI tests (with current challenges such as hallucinations and ethical bias) [23].

The perspective of a „minimalist" universalist deontological ethics of Kantian nature offers a solution not only for decision-making. The association of the idea of moral autonomy is grounded in the theoretical model of human reason in general, as developed in Kant's Critique of Pure

Reason [8]. The advantage is that, regardless of the structure and evolution of the whole and its AI components (symbolic AI, neural AI, etc.), the grounding of moral autonomy in the autonomy of human rational beings imposes a minimal, universal, and functional set that can be integrated into AI systems that can be tested [24][25][26].

Regarding the current crisis of the limited multimodal information (of human generated) online resources - again with a bias created by the presence of Internet-based ones, the same solution (from above) could address these limitations by "restricting at the source" any potential deviations related to biases generated by Internet-based resources. "Kantian-type filters" could "select" at the source data and information incompatible with the underlying model, and provide the fundamental attitude of knowledge evaluation. Recent developments in AI, particularly Agentic AI, raise the issue of the balance between "know-how you pay and know-what you get," given that the more one pursues increased autonomy and moves away from computational determinism, the greater the risk of generating results that are *ad-hoc*, unpredictable, or even hallucinatory (inaccurate or fictional). It is unlikely that these systems will approach AGI/ASI, given the divergence of current agentic AI solutions from the principles of existing tests for AGI (TT, Total TT [13], IQ, CAPTCHA).

If "traditional" computational ethical determinism can no longer be integrated into new models with Agentic AI, one solution might be the integration of the Kantian Attitude, including proposals of Kantian tests: even if it is not a "direct" filter focused on raw solutions (or perhaps it can still be?), it could still be present, potentially through a "backtracking" extension, in the formulation of the final answer, avoiding both incorrect responses and/or those incompatible with the principles of the test. Diagnosing the "current state" can be done relatively easy (failure to pass AGI/ASI tests). Here the concepts of the peaks of the two pyramids come into play, which can be opposed to the point of antithesis: the future MAL (Model Algorithm Library) cannot be placed over (Human) Wisdom; the presence of a Kantian test eliminates the additional risks that arise in its absence.

In addition to the concept of Kantian moral autonomy within the Kantian ethical test, operationalising the concept of teleology provides the seeds for the proposed *a posteriori* attitude evaluation. This concept can be present in the test through the general "orientation" of the process in Agentic AI, with the counterexample being the impossibility of constructing, in the current situation, a technological teleology. The concept of Kantian teleology integrated into a Kant-type test might not only hope to convert MAL into Wisdom but might even succeed in doing so as long as it is integrated from the lower levels of the pyramid: Clean Data/ Data level, Analytics-Indicators/ Information Level, Model and Algorithms /Knowledge level [27][28][29][30][31].

The claim of AGI as Singularity started raising opinions on the master-slave relationship in the human-robot developments. With reference to Hegel and Kant perspectives; a new set of open questions are expected: is GenAI currently in a slavery time within the human society? would the relationship change during the transition of GenAI in AGI? how can such relationship be measured using concepts and metrics since the transition requires intelligence engineering as the machine is learning to step towards artificial wisdom for machine learning know-what and know-how.

The issue of the "master-slave" relationship in the development of human-(ro)bot interactions becomes fundamental in the context of current trends to combine the two versions of AI (symbolic AI and neural networks): the dialectical metaphor of "master-slave" by Hegel tells us today that GenAI is in a period of "slavery" to human society (AI, through its outcomes, tends to become essential to human masters, who create and control – the "slave position"); but as AGI, will the direction of the relationship be maintained? Here, concepts such as moral autonomy, human dignity, and Kantian teleology can intervene to redesign the type of relationship originating from Hegel. For Kant, we have an asymmetrical relationship, human-AIG/ASI, which will be maintained in the form of an "assistance relationship"; for Hegel, the relationship is symmetrical. It is important to note that in the new configuration where Kant-type tests intervene by adjusting or integrating into symbolic AI + neural networks, the issue will not be about "competition," but rather about "assistance" of AI, with common goals and objectives, including "conform-oriented responses." The initial asymmetrical relationship (from Kant to Hegel), where the path is in one direction or both directions "human-robot," "master-slave" (vice-versa), will be replaced by one of "assistance." The Kantian concept of human dignity and its implications can be examined; it can be integrated to adjust the "moral status of AGI/ASI." Does the autonomy of AI lead to its "technological" dignity?

Kant can thus help to analyse whether the current human-AI relationship is merely instrumental, if AGI may one day surpass this paradigm, or if the problem can be framed differently. The answer lies in the nature of Kantian cognitive model of an essentially universal and deontological ethics, thus in the relationship type it establishes with the ethics of subjects (of this model). The use of this novel perspective will reconfigure this type of relationship [32][33][34][35]. Concluding, the main issue regards current inconsistent definitions and perceptions of AI: are we trying to humanise AI? Human Wisdom includes Learning, Reasoning, Understanding. We will name this relevant open set:

$$Attitude = \{Learning, Reasoning, Understanding\} \quad (1)$$

AI systems show already Learning and (partial) Reasoning abilities, but with different meanings and consequences of Understanding concept: the Understanding includes Conscientious attitude by default, including progress. Eq. (1) motivates and supports our proposal of Ethical AI meta-design: pure ethical reasoning supported by conscience [8].

IV. The Proposed Inclusive Framework for Ethical AI Meta-design and Evaluation

The current (AI) industrial revolution means enhancement of HW with AI from an utilitarian, business-oriented motivated, bottom-up learning framework, and not its replacement! If that is the case, do one need two different definitions of intelligence? Or parallel education systems? One may want for example to include the Hilbert perspective to AI systems in a similar way it is applied to fundament mathematics with 3 foci: correctness, completeness, and decidability, because we presume AI will move stronger in supporting human society in decision making tasks. This consistent approach could therefore show Godel's Collaboration (by *a priori* design) and Competition (by *a*

posteriori evaluation) in the next wave of AI towards AGI: all motivated by the Kantian ethical perspective.

AI as revolutionary digital technology started to elevate and re-define human-like reactions, but also humanity given current challenges and opportunities brought because of it, its dynamics and job pressures at top levels. Specialised Large Language Models algorithmic developments using data and information available online (publicly or with restricted access) are reshaping knowledge, cognition, innovation, human values. Progressive Large Action Models building on agentic AI are re-defining also actions and risks [10]. AI systems currently demonstrate with arguable evidence that can elevate and even re/define humanity given own current changing challenges and opportunities. This is interestingly aligned with the scientific origins of the Turing Test that focused on intelligence manifestations and their evaluations, on top of human(istic) perceptions of interactions.

However, the (presence identification of) Conscience makes the difference between artificial intelligence (AI) and human wisdom (HW) and is the bridge between logic and ethics in Kant's Critic of Pure Reason: *a priori* inclusivity by transcendental awareness, *a posteriori* reasoning in processes of inclusion of Kantian ethics in the design, production, testing, deployment, use and evaluation of AI systems.

```
# Function to check expert a priori Attitude //
essential for transcendental Kantian ethical
inclusion
def check_expert_attitude(role):
    if expert_available(role):
        return True
    else:
        request_expert(role)
        return False
# Step 1: Data Gathering and Analytics
def gather_data():
    if check_expert_attitude("data creators") and
check_expert_attitude("data analysts"):
        data_sources = ["text", "images", "audio",
"video", "etc"]
        raw_data = collect_data(data_sources)
        clean_data = preprocess_data(raw_data)
        return clean_data
# Step 2: Feature Engineering
def feature_engineering(clean_data):
    if check_expert_attitude("data engineers"):
        features = extract_features(clean_data)
        engineered_features =
transform_features(features)
        return engineered_features
# Step 3: Tokenization and Information
Representation
def tokenize_and_represent(engineered_features):
    if check_expert_attitude("data managers"):
        tokens =
tokenize_data(engineered_features)
        knowledge_representation =
represent_knowledge(tokens)
        return knowledge_representation
# Step 4: Training
def train_model(knowledge_representation):
    if check_expert_attitude("knowledge experts"):
        model = initialize_model()
        trained_model = train(model,
knowledge_representation)
        return trained_model
# Step 5: Testing
def test_model(trained_model):
    if check_expert_attitude("knowledge experts"):
        test_data = load_test_data()
        test_results =
evaluate_model(trained_model, test_data)
        return test_results
# Step 6: Deployment
def deploy_model(trained_model):
    if check_expert_attitude("knowledge experts"):
        deployment_environment =
setup_environment()
        deploy(trained_model,
deployment_environment)
# Step 7: Continuous Testing During Use
def continuous_testing(trained_model):
    if check_expert_attitude("knowledge experts"):
        while model_in_use():
            new_data = collect_new_data()
            test_results =
evaluate_model(trained_model, new_data)
            if performance_drops(test_results):
                retrain_model(trained_model,
new_data)
# Function of Evaluation of Performance, Costs and
Risks // essential to apply expert a posteriori
Attitude, essential for Kantian ethical inclusion
def
evaluate_performance_costs_&_risks(trained_model):
    if check_expert_attitude("evaluators"):
        performance_metrics =
calculate_performance_metrics(trained_model)
        costs_metrics =
assess_costs(trained_model)
        risks_metrics =
assess_risks(trained_model)
        return performance_metrics, costs_metrics,
risks_metrics

# Main Process
def main():
    if check_expert_attitude("data creators") and
check_expert_attitude("data analysts"):
        clean_data = gather_data()
    if check_expert_attitude("data engineers"):
        engineered_features =
feature_engineering(clean_data)
    if check_expert_attitude("data managers"):
        knowledge_representation =
tokenize_and_represent(engineered_features)
    if check_expert_attitude("knowledge experts"):
        trained_model =
train_model(knowledge_representation)
    if check_expert_attitude("knowledge experts"):
        test_results = test_model(trained_model)
    if check_expert_attitude("knowledge experts"):
        deploy_model(trained_model)
    if check_expert_attitude("knowledge experts"):
        continuous_testing(trained_model)
    if check_expert_attitude("evaluators"):
        performance_metrics, costs_metrics,
risks_metrics =
evaluate_performance_costs_&_risks(trained_model)
    report_results(performance_metrics,
costs_metrics, risks_metrics)
main()
```

Fig. 2. Pseudo-code of Ethical AI Meta-design Framework

A. Ethical AI Meta-design: framework and pseudo-code

Therefore, our pragmatic approach is to propose a collaborative framework for Ethical AI meta-design, with optimisation of Attitude elements' inclusion and contribution (Eq. 1) to motivate and integrate HW and AI in the creation of deontological Ethical AI systems. Within the traditional steps of GenAI production captured in the pseudocode (Fig. 2) we propose the following process, with *a priori* checks highlighted in the main GenAI creation and use, and the *a posteriori* evaluations as integrated parts of the AI systems lifetime. This meta-design expressed in pseudocode format outlines the key steps in developing a responsible Generative AI system, from data gathering to continuous evaluation of its maintenance and use. Each function(ality) represents a

required part of the process, ensuring the system remains responsibly and efficiently performant with clear governance steps from data/information gathering and analytics, feature engineering, knowledge representation, model training, testing, deployment, continuous testing during use, evaluation of performance, costs and risks, together with the proposed transcendental requirement of the human collaborators for conscientious ethical expertise.

Fig. 2 lists the meta-design in pseudo-code with *a priori* check for the necessity of Kantian Ethical Attitude in which "data creators" and "data analysts" are roles of "data scientists"; and "data" experts is used for simplicity of notation, being in fact "data & information" experts; the "knowledge experts" should again be refined in "knowledge scientists", "knowledge engineers" and "knowledge managers" in subsequent meta-design stages of relevance; the "evaluators" role is already above "knowledge" expertise and moves into the "intelligence" hierarchical domain as in the DIKW hierarchy. By introducing the combined meta-design of *a priori* attitude assessment + *a posteriori* evaluation in AI developments we break the challenges of AI complexities in a *divide et impera* approach of implicit and explicit ethical principles with Kantian perspective. The opportunity of Kantian ethics in moral AI design and evaluation is therefore twofold (Fig. 2) in the process of AGI creation and integration in human society: *a priori* transcendental by ethical attitude incorporation of Kantian morale (human agency) in the AI design loop – either by own intrinsic/default rules (as claimed by Anthropic AI) or an explicit learning of ethics by experts with reference to a library of Kantian Ethics; and *a posteriori* loop by experimental evaluation of AI systems by „AI Act" officers (human and artificial entities) by the use of established evaluation metrics and case studies.

B. DIKW Evaluation Metrics

Our proposed pragmatic meta-design of ethical AI within the AGI vision applies and builds on dianoetic approach to risk + virtue with argument for transhumanism [25][36]: humans must use technology proactively to augment and evolve as intelligent species. In this case, evolving towards new moral patterns for society and new norms build on approved machine ethics that may become our ethical adviser. Virtue as practical wisdom (phronesis) and habituation for understanding the transition in the risk hierarchy. In this sense, the current paper introduces the Entropy generalized views into the DIKW pyramid.

The main motivation is that, in the GenAI era, the AI systems are non-deterministic, chaotic, updateable and, in many respects, non-reliable. The reason is their architecture design and development, a mirror of continuously changing resources (of data, information and knowledge). The gaps in this entire process of fueling "deep neural networking" engines with uncontrollable quality of "oil" data and information create unknown quality outputs – many of high quality, some of lower or unknown quality. Such approaches are indeed losing even more transparency on AI models quality through globalization of transfer learning and agentic AI, with their commercial benefits at the cost of trust and risk. For the *a posteriori* component, we propose entropy for metrics of all DIKW dimensions of the AI space f discourse.

Entropy: the foundation stone starts from the well-known definition proposed by Claude Shannon in his Information Theory [37]: the study of quantification, storage and communication of information. Interestingly, Shannon used data and information interchangeably in his work. Entropy of a random variable X quantifies its uncertainty of potential states or unpredictability of possible outcomes. Higher entropy values show greater randomness (e.g. hallucinations); lower values signal deterministic behaviour (self-information). For its discrete form X with n values x_i of occurrences $p(x_i)$, the disorder of information for all possible outcomes $i = 1,..,n$ in a system is [37]:

$$H(X) = - \Sigma\, p(x_i)\, \log_2(p(x_i)) \quad (2)$$

Information Entropy concept is by default part of data analytics (in feature engineering using cleverly conditional probabilities) and AI models (in Decision Tree algorithms).

Moving up in the DIKW hierarchy, Knowledge Entropy describes levels of uncertainty, complexity, confusion, misinformation and disorder impacting know-what resources in general information systems [38] and LLMs [39]. Knowledge Entropy has potential in assessing scientific discoveries building from the philosophical view of the interplay between epistemology and rating of physical reality through probabilities. Given the complexity of knowledge parametric description for problems, the concept of knowledge entropy doesn't have a universally accepted formula, and no relevant quantified case study to date.

Moreover, Intelligence Entropy calculation is currently the challenge: the complexity of its features and their potential dependencies extrapolate the entropy equation to conditional probabilities combinatorically. Starting from (Eq.1) one can argue that Attitude's variables (Learning, Reasoning and Understanding) show the clear border between AI and HW in the fact that there is independence of Conscience for AI systems, and conditional dependencies by Conscience for HW, while both have already quantifiable outputs from (Performance, Costs, Risks).

The need of Intelligence Entropy is justified by the current AI's non-determinist, influenceable character of GenAI evolutions. This is supported by Godel Theorem in the context of AI Governance: a complex system is either incomplete or inconsistent. Relevant examples are TT and IQ tests [40] although they are dynamically evolving, since there is still a debate of such tests primarily measuring reasoning and cognitive abilities, while AI systems are catching up fast at least in areas of pattern recognition, logical reasoning and theorem demonstration and applications.

V. Case Study: Research Publications and GenAI

The case study explores the complexities of the inter-sectorial presence of AI and HW in the process of creation and authorship, scientific review and editorial management of scientific publications. In the past years there are more pressures [5] from funders, institutions and publishers for increases in publication quantities, their metrics (citations, impact factors, rankings) at the potential costs of authorship and quality approaches [6]. The Data & Information Entropies [37] applied to examples from Retraction Watch Data prove an interesting domain for our Kantian ethical meta-design with entropy evaluations. The motivation is that this domain is biased and prone to fallacies. While in most cases all actors expect and provide explicitly checks against GenAI-generated texts with GenAI tools, one can build the confusion matrix as such: True Negatives (TNs) are human-authored, no GenAI nor plagiarism cases. Papers are characterised by the main

topic, and faulty papers could be clean, or have minor/major text plagiarism, data, image, citation issues or full plagiarism status. Trained GenAI could be used both, for creating (and hiding?) plagiarism, as well as for detecting plagiarism. Of course their design and training vary with diverse outputs: clean papers to get through, and various academic misconduct cases to be detected. In other terms TN = good papers, and TP bad papers. Let's argue on "normal" statistical scenarios.

Based on a review of 1000 papers for a given journal, let's say these frequencies were found: Clean papers: 850 cases (p = 0.850), Minor text recycling: 94 cases (p = 0.094), Major text plagiarism: 15 cases (p = 0.015), Data plagiarism: 36 cases (p = 0.030), Complete paper plagiarism: 5 cases (p = 0.005). Plagiarism Information Entropy is:

$$H = -(0.850 \times \log_2(0.850) + 0.094 \times \log_2(0.094) + 0.015 \times \log_2(0.015) + 0.036 \times \log_2(0.036) + 0.005 \times \log_2(0.005))$$

$$H = -(-0.199 - 0.332 - 0.090 - 0.151 - 0.038) = 0.81 \text{ bits}$$

The relatively low entropy (maximum would be $\log_2(5) \approx 2.32$ bits) reflects the highly skewed nature of the distribution. This suggests that the system is fairly predictable, with most papers being clean.

Let's consider now from a RetractionWatch [6] fresh record perspective (figures are relative) among papers retracted for plagiarism: Text plagiarism: ~42% (p = 0.420); Image/figure plagiarism: ~28% (p = 0.280); Self-plagiarism: ~18% (p = 0.180); Data plagiarism: ~8% (p = 0.080); Reference/citation plagiarism: ~4% (p = 0.040). The detection entropy is:

$$H = -(0.420 \times \log_2(0.420) + 0.280 \times \log_2(0.280) + 0.180 \times \log_2(0.180) + 0.080 \times \log_2(0.080) + 0.040 \times \log_2(0.040))$$

$$H = -(-0.526 - 0.514 - 0.445 - 0.292 - 0.186) = 1.963 \text{ bits}$$

The entropy of 1.963 bits (compared to our previous hypothetical example of 0.81) suggests more evenly distributed types of plagiarism among detected cases, with greater uncertainty in predicting what type of plagiarism might occur, and more complex detection challenges.

However, while acknowledging RetractionWatch [6] numbers based on papers retracted, the decision to do so could be based on fluid data, that increased heavily over time [41] including difference in text, image and data generation, perplexity, susceptibility, validity measurements. But there are FNs landings, and FPs dangers [42] of unfairly plagiarism accusations by GenAI detection tools because of potential algorithmic bias, data imbalance, statistical generalisations (between topics, journals and authors), with strong examples for text written by non-native English speakers, autistic style, topic, and the GenAI humanisation race to pass TT, IQ tests. The corroborated danger: a TP or FP retracted paper creates waves of impact in subsequent papers citing them. This is a corresponding knowledge entropy evaluation with impact. Our proposal for collaborative meta-design stands, involving collaboratively automated tools and human experts in both a priori involvement of consciously ethical experts, and a posteriori (statistical) evaluation of tools.

VI. Conclusions and Future Work

The proposed meta-design framework and use of entropy metrics acknowledge that AI systems, unlike traditional software, have the potential for non-deterministic, continuous adaptation and learning with a pragmatic vision [43]. This means that the 'use time' becomes a critical phase where the system can evolve beyond its initial design constraints through the *a-priori* trained integration of ethics along logical AI design and development, and *a-posteriori* evaluation during use and continuous training by: 1) dynamic adaptation of AI know-what and know-how based on interaction patterns; 2) trustful evolution of knowledge representations and decision-making processes; 3) continuously evaluated refinement of learning, reasoning and understanding for all; 4) emergence of new capabilities through human-AI collaboration for all (as captured in Fig. 2.) The proposed Kantian ethics meta-design approach acknowledges, creates and manages a more fluid yet trustful bridge between design and operational phases, where AI system(s) and human designers, developers, evaluators and users become both product of design and active co-participants in evolution by co-creation.

The next stages of our pragmatic still theoretical contributions are to design experimental evidence gathering within the area of metrics with detailed implementation statistics, with case studies on Knowledge and Intelligence Entropies to support the theoretical arguments presented hereby. Such case studies will also allow building the framework of the first steps of practical implementation of the proposed inclusive Framework for Ethical AI Meta-Design with co-creative checks between the human and the machine collaborators. In this way this contribution will progress further on ways to integrate it by means of explicit, preferably explainable interactive feedback between the human and the machine for continuous learning. Real life examples within computational developments will then be realistically scheduled, preferably in open source GenAI versions. The planned experimental evidence will be firstly in IQ and TT with focus on co-creative challenges, connected to ethics, aesthetics and logic, therefore with applications in arts, scientific authoring and education. Consequently, detailed explanations within the case studies will add details on proposed metrics, methodologies and framework.

The proposed ethical meta-design framework claims a similar integration with DNA-RNA roles in life, to replace a potential "master-slave" relationship. While information for proteins with genetic roles is contained in DNA, the encoding (i.e. development) in amino acids and the transfer (i.e. deployment) to ribosomes to make proteins is the RNA role. We advocate through Kantian ethics a DNA-RNA future for HW-AI.

References

[1] P.P. Groumpos, "A critical historical and scientific overview of all industrial revolutions", IFAC-PapersOnLine, vol. 54/13, pp. 464-471, 2021. DOI: 10.1016/j.ifacol.2021.10.492

[2] M.A.Peters, P. Jandrić, B.J. Green, "The DIKW Model in the Age of AI". *Postdig Sci & Ed,* 2024. DOI: 10.1007/s42438-024-00462-8

[3] M. Frické "The knowledge pyramid: a critique of the DIKW hierarchy" J.Inf.Sci,35(2), pp. 131-142, 2009. DOI: 10.1177/0165551508094050

[4] D. Weinberger, "The Problem with the Data-Information-Knowledge-Wisdom Hierarchy". Harvard Business Review, 2010. [Online]. Available: https://hbr.org/2010/02/data-is-to-info-as-info-is-not

[5] T. Kron, "Scientific Publications Face Credibility Crisis" Medscape. 2025[Online]Available:https://www.medscape.com/viewarticle/scientific-publications-face-credibility-crisis-2025a10000fb

[6] Retraction Watch – Tracking retractions as a window into the scientific process, 2025 [Online] Available: https://retractionwatch.com/

[7] S.V. Ahamed, "Next Generation Knowledge Machines", Elsevier, pp. 233-263, 2014. https://doi.org/10.1016/C2012-0-06125-X

[8] Imm. Kant, Critique of Pure Reason. Cambridge University Press. 1781. Edited by P Guyer, AW Wood, 2013.

[9] H. Kim, and D. Schönecker (eds.) Kant and Artificial Intelligence, Berlin, Boston: De Gruyter, 2022. DOI10.1515/9783110706611

[10] EU. "Shaping Europe's digital future". 2025. [Online] Available https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[11] P. Rowady. DIKW hierarchy w Automation Equivalent – Alphacution Research Conservatory. 2024 [Online] Available https://alphacution.com/the-decay-of-knowledge-in-douchebagistan/

[12] A. Turing, "Computing Machinery and Intelligence". Mind 49: pp. 433-460. 1950

[13] P. Schweizer, "The Truly Total Turing Test*". Minds and Machines 8, pp. 263–272, 1998. https://doi.org/10.1023/A:1008229619541

[14] A. Rogoyski (2024) AI has a stupid secret: we're still not sure how to test for human levels of intelligence. The Conversation, 4 Oct 2024. [Online] Available: https://theconversation.com/ai-has-a-stupid-secret-were-still-not-sure-how-to-test-for-human-levels-of-intelligence-240469

[15] F. Chollet. "On the Measure of Intelligence" arXiv:1911.01547v2. 2019. [Online]. Available: https://arxiv.org/pdf/1911.01547

[16] N. Dinh, L. Ogiela, Human-artificial intelligence approaches for secure analysis in CAPTCHA codes. EURASIP J. on Info. Security 2022, 8, 2022. https://doi.org/10.1186/s13635-022-00134-9

[17] M.T. Bennett, E. Perrier, "An AI system has reached human level on a test for 'general intelligence'. Here's what that means" The Conversation, 24 December 2024. [Online] Available: https://theconversation.com/an-ai-system-has-reached-human-level-on-a-test-for-general-intelligence-heres-what-that-means-246529

[18] F. Chollet. "Abstraction and Reasoning Corpus for Artificial General Intelligence (ARC-AGI)", 2025. [Online] Available https://github.com/fchollet/ARC-AGI

[19] ARC Prize 2025 [Online] Available https://arcprize.org/arc

[20] R. Manna, R. Nath, "Kantian Moral Agency and the Ethics of Artificial Intelligence", Problemos, 100, pp.139–151, 2021. doi:10.15388/Problemos.100.11

[21] M.A. Drăghici, "Kantian Problems in Contemporary Epistemology (Probleme kantiene în epistemologia contemporană)". Editura Academiei Române, București, 2016, pp. 157, 169.

[22] I. Neagu, "On the Relevance of the Kantian Perspective in the Era of Generative Artificial Intelligence: Ethical and Technical Challenges", Journal of Philosophy, tome LXXI – 2024, Issue 3, 2024, pp. 1-10.

[23] S. Farquhar, J. Kossen, L. Kuhn, et al. "Detecting hallucinations in large language models using semantic entropy". Nature 630, 625–630, 2024. https://doi.org/10.1038/s41586-024-07421-0

[24] T. Schlicht, "Minds, Brains, and Deep Learning: The Development of Cognitive Science Through the Lens of Kant's Approach to Cognition," Kant and AI, De Gruyter, 2022, pp. 3–38.

[25] M. Constantinescu, C. Voinea, R. Uszkai, et al. "Understanding responsibility in Responsible AI. Dianoetic virtues and the hard problem of context". Ethics Inf Technol 23, 803–814, 2021. https://doi.org/10.1007/s10676-021-09616-9

[26] L. Benossi, S. Bernecker, "A Kantian Perspective on Robot Ethics," Kant and AI, De Gruyter, 2022, pp. 147–168.

[27] R. Evans, "The Apperception Engine," Kant and AI, De Gruyter, 2022, pp. 39-103.

[28] S. Baiasu, "The Challenge of (Self)Consciousness: Kant, AI and Sense-Making," Kant and AI, De Gruyter, 2022, pp. 105–128.

[29] E.E. Schmidt, "Kant on Trolleys and Autonomous Driving," Kant and AI, De Gruyter, 2022, pp. 189–222.

[30] D.J. Chalmers, "The Conscious Mind. In Search of a Fundamental Theory," Oxford: Oxford University Press, 1996.

[31] D. Schönecker, "Can Practical Reason Be Artificial," Journal of Artificial Intelligence and Humanities, 2, pp. 67–91. ISSN: 2951-388X

[32] H. Kim, "Tracing the Origins of Artificial Intelligence: A Kantian Response to McCarthy's Call for Philosophical Help," Kant and AI, De Gruyter, 2022, pp. 129–146.

[33] R.B. Louden, "Kant's Impure Ethics From Rational Beings to Human Beings", Oxford Univ. Press, 2000. DOI:10.1093/oso/9780195130416.001.0001

[34] A. Thomas Wright, "Rightful Machines," Kant and AI, De Gruyter, 2022, pp. 223–238.

[35] C. Dierksmeier, "Partners, Not Parts. Enhanced Autonomy Through Artificial Intelligence? A Kantian Perspective," Kant and AI, De Gruyter, 2024, p. 239–280.

[36] S. Vallor, "The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking (New York, 2024; online edn, Oxford Academic, 30 Apr. 2024), https://doi.org/10.1093/oso/9780197759066.001.0001

[37] C. E. Shannon, "A mathematical theory of communication," in The Bell System Technical Journal, vol. 27, no. 3, pp. 379-423, July 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x

[38] HJ. Kelley. "Entropy of Knowledge". Philosophy of Science. 1969; 36(2):178-196. DOI: https://doi.org/10.1086/288244

[39] J. Kim, H. Lee, H. Cho, J. Jang, H. Hwang, S. Won, Y. Ahn, D. Lee, M. Seo, "Knowledge Entropy Decay during Language Model Pretraining Hinders New Knowledge Acquisition" (2024) Preprint https://doi.org/10.48550/arXiv.2410.01380

[40] PH. Schönemann "Do IQ tests really measure intelligence?" Behavioral and Brain Sciences, 6/2, 1983, pp. 311–313 https://doi.org/10.1017/S0140525X00016125

[41] L. Giray, K. Sevnarayan, F.R. Madiseh: "Beyond Policing: AI Writing Detection Tools, Trust, Academic Integrity, and Their Implications for College Writing", Internet Reference Services Quarterly, 2025. DOI: 10.1080/10875301.2024.2437174

[42] L. Giray. "The Problem with False Positives: AI Detection Unfairly Accuses Scholars of AI Plagiarism". The Serials Librarian, 2024. 1–9. https://doi.org/10.1080/0361526X.2024.2433256

[43] D.S. Watson, J. Mökander, L. Floridi, L. Competing narratives in AI ethics: a defense of sociotechnical pragmatism. AI & Soc., 2024. https://doi.org/10.1007/s00146-024-02128-2