Analele Universității din București. Filosofie
Vol. LXXIII – Nr. 2, 2024, p. 57-74
ISSN 0068-3175; e-ISSN 2537-4044
DOI: 10.62229/aubpslxxiii/2_24/5

Sorin CREȚU[1]

THE "THREE-SYSTEM VIEW" MODEL. A POSSIBLE SOLUTION SOURCED IN "PRE-JUDICATIVE HERMENEUTICS"

Abstract. The current paper, anchored in the discussions surrounding "*Linda Problem*", "*model-free*" and "*model-based*" learning strategies along with Kahneman's "*two-system view*" cognitive model, and Gross's considerations on "*salience processing*" and "*atypical attribution of salience*", aims to present a potential solution originating in *Pre-Judicative Hermeneutics*: Cernica's "*three-system view*" model. Both Kahneman's and Cernica's models embrace Herbert Simon's paradigm of "*bounded rationality*" and serve to broaden its reach. While Kahneman's "*System1*" and "*System2*" model elaborates on the "*perception*" – "*intuition*" – "*reasoning*" cycle, Cernica's model, founded upon and expending the "*non-judicative*" – "*pre-judicative*" – "*judicative*" cycle, is intended, I suggest, to augment and advance the former, from and with a phenomenological and hermeneutical perspective. Core concepts as "*pre-judicative memory*", "*de-constitution*" "*interpretation*", "*pre-judicative circularity*", "*act of pre-judging*", "*non-judicative experience* "and "*existential judgment*" are central to Cernica's innovative approach. These ideas bring the "*existential subject*" ("*operant subject*") in focus, establishing a framework to account for how the cognitive biases could be mitigated across a wide range of domains, thus enabling the elevation of our consciousness and existence through the fundamental attributes of "*reality*", "*identity*", "*authenticity*" and "*veracity*" supporting them. Furthermore, as reiterated in this paper and elaborated on in a more extensive discussion in a previous work, the enhancements proposed through Cernica's "*three-system view*" model may also prove instrumental in emerging areas of AI development and the pursuit of AGI. These include Reinforcement Learning (RL) strategies, AI Safety and AI Alignment research. The model offers valuable contributions to a deeper understanding of debiasing processes, the development of more robust

[1] PhD, Doctoral School of Philosophy, Faculty of Philosophy, University of Bucharest, <soleliom@gmail.com>. Member of the SHPjM Research Group (*Studies in Pre-Judicative Hermeneutics and Meontology*), coordinated by Prof. Univ. Dr. Viorel Cernica, University of Bucharest. SHPjM Research Group is part of Research Center of the History of Philosophical Ideas (CCIIF).

EDITURA UNIVERSITĂȚII DIN BUCUREȘTI
BUCHAREST UNIVERSITY PRESS

debiasing methods, and a more nuanced understanding and alignment of "*deliberative human preferences*" and "*human value functions*" within advanced AI models.

Keywords: *non-judicative, pre-judicative, judicative; operant subject, logical subject, reality, identity, authenticity, veracity, pre-judicative memory, de-constitution, re-interpretation, pre-judicative circularity, cognitive bias, deliberative human preferences, bounded rationality, shared intersubjective space*

The perplexing findings related to the answers to "*Linda Problem*" (Kahneman 2002, 468), particularly reflective of massive "*attribute-substitution*" (469), in the context of observed "*accessibility*" challenges, uncovering the permeation of the representational content of "*System2*" with pre-judgmental elements from Kahneman's "*System1*" (451)[2] (*e.g.*, intuitions, intuitive "*preferences*"[3] based on "*impressions*" on the attributes of "perceptual and cognitive *objects*"[4]) that escaped appropriate monitoring and deliberative qualification, as well as the surprising observations on the topic of "*salience attribution*" and "*salience processing*" that exposed us to the reinterpretation, with the supportive evidence of neuroscience, of the notion of "*importance*", topic thus expended into "*atypical attribution of salience*" (Gross *et al.* 2024) characterizing creative processes and judgments, bring us to an inflection point. It is the point at which, I suggest, we should look for an out-of-the-box alternative, with the thought that this could bridge the aforementioned difficulties towards clarifying opportunities.

To open this discussion, a few daring questions should be asked beforehand.

- In what ways our "*pre-judicative memory*" can play the fundamental "*de-constitutive*" role in rebalancing our judgments towards "*value-driven*" perspectives?
- Why our aesthetic judgments, also scoped in this paper, and along with them, I suggest, other particular judgments highly cherished

[2] See "*Figure 1*", mapping the process flow between "*Perception*", "*System1*" and "*System2*".

[3] The notion of "*preferences*" should be also observed in connection with the Reinforcement Learning (RL) framework from Artificial Intelligence (AI) and in the context of "*human preferences*", which are representative of the human judgments; this discussion was extensively elaborated in my paper, Crețu 2025 (forthcoming).

[4] Previously discussed in Crețu 2025.

throughout the human domain (*e.g.*, affective, spiritual judgments) are so profoundly impacted by the "*non-judicative*" – "*pre-judicative*" – "*judicative*" cycle?

- At what point the "*operant subject*" supersedes the "*logical subject*" such that our way of experiencing the world can upgrade our existence with the attributes of "*reality*", "*identity*", "*authenticity*" and "*veracity*"?

Such a possible alternative is the solution model proposed by Viorel Cernica, within the framework of "*pre-judicative*[5] *hermeneutics*" (see Cernica 2019, 185; 249). As previously emphasized (Crețu 2024), the appeal to the hermeneutics' mechanisms of judgment evaluation and cognitive bias de-constitution could be a promising avenue, due to its dedicated interpretative capabilities (also applicable when we approach AI Alignment research) when we talk about "*evaluative attitudes*" (in terms of "*desires*"; particularly "*instrumental desires*", and "*higher-order desires*", the former leading to the latter, at least in some instance), "*human preferences*" (as "*comparative evaluations*", particularly in their deliberative form), intentions (alternatives-based selections) and "*evaluative beliefs*", along with considering the cognitive "*salience*" elements, all these aspects pointing to both "*value*" and "*choice*" differentiation and consolidation and "*value-driven*" behaviors and actions.

The context of this solution model revolves around multiple core concepts reflective of their domain specific acts and structures that are linked at the at aspectual (process) level and at the role (functional) level, being tied into the "*operant subject*" ("*in person*", differentiated ontologically from the "*logical subject*", meaning that the "*operant subject*" and "*logical subject*" stand in an "*ontological difference*" and I, would suggest, tension)

[5] As a clarification pertaining to the term "*judicative*" captured in the concepts of "*judicative*", "*pre-judicative*" and "*non-judicative*" employed throughout Viorel Cernica's model: the term is not related in any way to the "*judicative*" attribute used within the judicial framework. Furthermore, the term "*prejudice*" that we will refer to as part of this model is not intended to be related to "*an affective component*" (*e.g.*, "*discriminatory behavior*"), but to "*a cognitive component*", cognitive bias-related; Refer to *APA Dictionary of Psychology*, from American Psychological Association, available online at < https://dictionary.apa.org/ > last time accessed on March 29, 2025, under the noun "*prejudice*".

(Cernica 2023, 267; 298)[6] with its experience and existence – refined under its existentialist attributes of "*reality*", "*identity*", "*authenticity*" (276-277) and "*veracity*"[7] (human, and which could be extended to AI agents): "*judicative*", "*pre-judicative*", "*non-judicative*", "*non-cognitive*"[8] (Cernica 2019, 185; 249). Once again, the term "*judicative*" refers to "*judgements*", being applied under their types. In this tripartite configuration of the model, the "*pre-judicative*" (Cernica 2019, 205) domain mediates between the "*judicative*" domain and "*non-judicative*" domain, based on the particular processes pertaining to each one of them (cognitive processes being associated with the "*judicative*" domain, while intuitive and emotional scopes, falling under "*non-cognitive*" processes underlying the "*non-judicative*" domain). We can already notice the parallel with Kahneman's aforementioned "*System1*" and "*System2*" cognitive models and with the "*model-free*" and "*model-based*" computational learning strategies, both extensively discussed in a previous paper as well[9].

These three domains stated earlier are dynamically interacting, based on particular operations, out of which the central ones are the "*operations of de-constitution*", the "*operations of interpretation*" and the "*intentional operations*". Naturally, the "*operations of interpretation*" rely on the "*operations of de-constitution*" of judgments from the "*judicative*" domain. Suffice to specify that the judgments referred to within the "*judicative*" domain are based on "*judicative structures*" and explicit reasoning processes. These "*judicative structures*" are integrated at cognitive level (akin to

[6] See also Cernica 2023, 276-277 and 274, for the "*ontological difference*".

[7] For a more extensive discussion of the four existentialist attributes – and, more precisely, the "*subject's identity hypostases*" – "*reality*", "*identity*", "*authenticity*" and "*veracity*" – I refer to Cernica's "Pre-Judicative Memory and De-Constituting of the Prejudices [Logic-formal and hermeneutic description of the existential subject's experience]", Chapter 2.31 "Objectual fulfilment of the subject (reality, identity, authenticity, veracity) and memory" (forthcoming publication).

[8] "Non-judicative thought according to quantity – the only criterion that can work, for now – covers a vast 'logic' and 'extra-logic' space, the latter being 'logicized': a) different logical forms of judgment (…), b) pre-judicative entities, c) pre-logical elements of a psychological nature, partially 'logicized': experiences, facts, especially those of the type of prejudices and pre-understandings (…)" (Cernica 2019, 69).

[9] In reference to the chapters "Predictive Human RL Strategies: 'Model-Free' and 'Model-Based'" (in particular, 70-75) and "Kahneman's 'System1' and 'System2' Cognitive Models" (in particular, 76-91) see Crețu 2025 (forthcoming).

"*System1*"), but they are functionally differentiated and instantiated as "*formal*" structures ("*subject-predicate*" type), "*alethic*" structures ("*verb-time*" relationship type), "*referential*" structures ("*expression-reference*" relationship type, where the "*expression*" is symbolically captured through language at the propositional level and the reference is a "*state of affairs*") and "*existential*" structures ("*existential subject*" – its "*pre-judicative memory*"; see Cernica 2023, 267; 298[10]). Thus, a cognitive judgment could be "*analysed*" either based on its "*structural makeup*", consisting of the previously mentioned four cognitive instantiations, or on its "*pre-judicative makeup*" (Cernica 2023, 273). The "*pre-judicative makeup*", to translate, paraphrasing the author, could be analysed within the framework of "*pre-judicative analytics*", by examining the "*de-constitution conditions of pre-judgments*" (273).

This "*pre-judicative makeup*" of judgments contains the "*cognitive judgment*", the "*evaluative judgment*" (applied on the "*cognitive judgment*", which becomes subject of evaluation, function of truth conditions) and the "*existential judgment*" (of the type "*I exist*", in which the "*I*" is the "*operant subject*" performing the "*judicative*" act, paraphrasing the author) (273).

We observe how the cognitive judgement "*structural makeup*" intersects the "*pre-judicative makeup*" at the level of the "*existential subject*" and to what extent the "*pre-judicative makeup*" intersects the cognitive domain at the level of an embedded "*cognitive judgment*", then differentiating from the cognitive domain through an evaluative and then existential slide towards the "*operant subject*", who "*performed*" the act of judgement in the first place (276). This slide towards the subject, engages directly the "*non-judicative*" domain, from which, on the one side, the aspectual part of it (at process-level) will reverberate back, towards the "*pre-judicative*" domain, infusing perceptions into pre-judgments and, on the other side, the representational part of it will reverberate back, towards the "*judicative domain*", in the form of representational content (role-wise, including the functional role, but not

[10] "The operant subject, flesh and blood person, and his pre-judicative memory, *e.g.*, the mass of traces left by his judicative experiences, his 'knowledge', now activated)" (Cernica 2023, 275). See also Cernica 2023, 271; 272; 275.

limited to it) transferred throughout the judgments' structure, permeating them starting from the "*existential*" structures, into "*referential*" structures and, further on, spreading through the "*judicative*" core, at "alethic" (unfolding in time a state or an action) or "*formal*" (reasoning) levels (engaging mostly our deductive reasoning through "*declarative knowledge*" and inductive and causal reasoning through "*procedural knowledge*", borderline with adductive and analogical reasoning, which already involve our "*symbolic knowledge*" – to extrapolate from Ben Goertzel's AGI cognitive infrastructure I was discussing in Crețu 2023, 124; 147).

Once again, we can see how aspectual and role levels are intertwined and, although there are content barriers (representational and conceptual in nature) between "*perception*" and "*intuition*" and aspectual barriers between "*intuition*" and "*reasoning*", the "*non-judicative*" and "*pre-judicative*" domains continue to interact with the cognitive domain, eluding the barriers (when the barrier is aspectual, the bridge is the content and when the barrier is the content, the crossing is aspectual). This incursion into the continuous cycle "*non-judicative*" – "*pre-judicative*" – "*judicative*" displays the convoluted path of cognitive biases, using the heuristic transfers between our perceptions and intentions, which is then taken over by the cognitive transfers between our intuitions and our reasoning, where the corrective filtering intervenes, and the deliberative processes occur[11]. While we are re-emphasizing Kahneman's "*two-system view*" (450) ("*System1*" and "*System2*", already manifesting throughout the "*perception*" – "*intuition*" – "*reasoning*" cycle), we are placing it in the context of a "*three-system view*" as described in Cernica's model ("*non-judicative*" – "*pre-judicative*" – "*judicative*" cycle), in which its specific continuous cycle is dominated by "*pre-judicative circularity*".

What the "*intuition*" and the "*reasoning*" continue to have in common (distinct from "*perception*") in Kahneman's "*two-system view*" is the passage of time and the fact that they "*can be evoked by language*" (451), which are maintained in Cernica's "*three-system view*" model at the "*judicative*" level, both through the "*alethic*" structures and the "*formal*" and "*referential*" structures.

[11] Re-emphasizing Kahneman's "*perception*" – "*intuition*" – "*reasoning*" linkages and differentiations; see Kahneman 2002, 450-451, including "*Figure1*", 452.

This "*three-system view*" allows the phenomenological and hermeneutical elements to take part in an integrative solution that seemed to have been mostly sought within the spheres of cognitive science. What makes possible this expansion into a "*three-system view*" is a previous and ongoing paradigm shift towards "*bounded rationality*" (Crețu 2023a, 171-178), as proposed by Herbert Simon and adopted by Kahneman in his "*maps of bounded rationality*", in which he is reiterating the "*two-system view*" human cognitive model, with heuristic foundations and exposing the cognitive biases processes.

Moreover, the "*bounded rationality*" paradigm had profound repercussions on the long-lasting, widely accepted and mostly unquestioned, by that time, "*unbounded rationality*" view, that used to proclaim the powers of human reason as "*universal*" and "*abstract*", with an "*absolute*" defying hint. The "*embodied realism*" that flourished in the wake of the new paradigm, particularly at the "*School of Berkeley*" through Lakoff's "*embodied mind theory*", brings forward through the wide-open avenue of human "*embodiment*" and "*embodied cognition*" (expended to and combined further on with "*embodied simulation*" perspectives and detailed through theories like "*shared manifold hypothesis*" or "*the shared manifold of intersubjectivity*" (Crețu 2023a, 93; 98), developed by Vittorio Gallese, at the "*School of Parma*"), both the structural and functional limitations of our reasoning, by "*closing two major gaps: the one between perception and conception and the one between subject and object, in which misleading objectivity ideas originated*" (Crețu 2023a, 191-192). This is the long-waited return of the "*operant subject*" (and, furthermore, of the "*first-person*" view) at the level of "*judicative*" reasoning constitution, through "*existential*" structures already implicitly present (Cernica 2023, 275-276), as part of a broader answer to deeper questions related to the naturalization of human intentionality, the nature of human integration and interaction with the world(within a "*multimodal intentional shared space*" or "*shared intersubjective space*", which constitute, foundationally, the human "*we-centric space*" as proposed by Gallese (2003, 525)[12], and the nature of consciousness.

[12] See also my paper where I discussed these concepts (Gallese 2003, 525 and Crețu 2023b, 93, 97-99), "The Advent of AGI".

What was needed as part of the solution was to bring the "*operant subject*" into our line of sight, a theoretical shift that Cernica's "*three-system view*" model permitted, at the level of "*existential*" structures within the "*judicative*" domain, where the "*existential subject*" is directly linked to its "*pre-judicative memory*" that enters the "*pre-judicative makeup*" of judgment through the "*existential judgment*" component (Cernica 2023, 274)[13]. Such dynamic attempts bridge the content gaps with subject's "*perception*" that ultimately reinforce our biases, by placing them in a conceptual representational space.

These aspects have also a direct impact on AI development and transformation towards AGI, including RL(Reinforcement Learning), IRL (Inverse Reinforcement Learning), the balance between "*reward functions*", "*objective functions*" and "*value functions*", and AI Alignment research, which explores, as we discussed in an earlier paper[14], the limitations and the prioritization of "*human reward functions*" versus "*human value functions*" (for the latter, in terms of "*human preferences*", and, as I suggested, "*deliberative human preferences*") as "*target for alignment*"[15].

I highlighted these points to show the directions related to the "*what*", in the question "*what is the solution we are looking for?*". It seems that we already know a lot about the "*what*". As far as the "*why*", by now, we also have a rather clear picture.

However, we don't clearly know the "*how*". The "*how*" posses significant challenges and it is the particularity of this solution that it could be providing a possible answer to the "*how*" in the question: How can we resolve cognitive biases, while remaining aligned with our heuristic, "*bounded rationality*" mechanisms and within an "*embodied realist*" framework, and also include in the process the "*operant subject*", such that we are able to account for the experiential aspects, and, with a target, at the "*first-person*" level, throughout our intentional cognitive processes, in order to set clear and transparent expectations regarding

[13] Here the author addresses the question, opening up a possible direction for further investigation: "But wouldn't be a possible connection between this judgment (existential) and the existential structure within the structural composition of the judgment (existential subject – its pre-judicative memory)?"

[14] See Crețu 2025 (forthcoming).

[15] These aspects have been extensively discussed in Crețu 2025.

our decision-making, our choices, our actions, our goals? The difficulties the AI Alignment research are pointing to are genuine. The "*human in the loop*" perspective requires that humans clarify their own domain, such that AI can align with, under conditions that both humans and AI can fully and deeply understand and follow. If we ever say, "*We didn't know better*", it will be because we really didn't know "*How*". Many times, behind the question "*What was missing?*" lies the real question we need to confront: "*How was it lost*?"

At this point, we can already glimpse the possible solution we referred to earlier, but before expending on that, we have to explicitly showcase the main elements along with their counterpoints, linking "*judgement*" to "*experience*" to "*act*" and to "*operation*": along with "*judgement*" and "*pre-judgment*"[16], the "*non-judicative experience (which is opposed to the judicative one)*", "*the act of pre-judging (which is opposed to the act of judging)*" and "*the operation of interpretation (which is opposed to the intentional operations)*" (Cernica 2019, 185). This "*judgement-experience-act-operation*" overarching organization of the model is an indication of its active and agentive capabilities, that will be confirmed by the appeal to "*pre-judicative memory*". Within this structure, three other concepts are playing a core role: "*de-constitution*", the "*pre-judicative circularity*" and the concept of "*interpretation*" (188)[17]. While the "*pre-judicative circularity*" that we previously mentioned is the "*movement of acts of consciousness between the judgment and the result of the non-judicative experience, which has the meaning of a 'reduction' of the two to a middle term: the pre-judicative*" (205), the "*interpretation*" is sourced in the "*experience of the non-judicative, whose content is processed in such a way that, taking the form of judgment, it becomes 'knowledge' and the object of communication*" (205).

The "*interpretation*" is applied to the "*judicative form*" but taking into account "*(…) the existential judgment from the pre-judicative composition of a judicative form* (…)" (Cernica 2023, 274). The "*operant subject*" is not

[16] "The pre-judgement is the formal outcome of the act of pre-judging, meaning the act of appropriation of an (cognitive) object through what another (person) or myself (…) has already thought (judged) about it, and then sent it to *me* in its standard (judicative) form of comprehension." (Cernica 2019, 189)

[17] We will not dive at this time into other connected concepts like "*negative predication*" and "*existential time*", concepts also mentioned by the author.

"*obviously*" entering the "*structural analysis of the judicative form*", but he is instrumental in its "*constitution*", as the judgment itself is "*his work*" (276), as being the inner, active part of his own "*existential*" segments of judgments through its "*pre-judicative memory*", throughout the "*judicative*" reasoning process.

The "*interpretation*", by the means of the "*pre-judicative memory*", is the active "*operant subject*'s" intervention in the judgement process from a metacognitive reasoning perspective (which, as we know, like the interpersonal/ intersubjective reasoning, relies on "*attentional knowledge*"), process that, himself, owns. This intervention is performed by consciously suspending a judgement and orienting the attentional resources towards following back the "*traces left by his judicative experiences*" (275) (to reiterate), which were previously integrated, and progressively lesser subjected to a questioning and qualification effort, in the subject's body of knowledge at the judicative, reasoning level, assumed to represent the "*logical subject*" point of view. These "*traces*" identified as "*active*" in the present "*pre-judicative memory*" are then exposed to an iterative process of "*de-constitution*", followed by "*interpretation*" – meaning here the "*leap from logical-formal to existential, a leap made possible by analysis*" of pre-judgements (279) and "*re-interpretation*" of those judgments, thus purging from them the cognitive bias elements, which were covertly woven within those "*traces*". As we observe the "*pre-judicative memory*" process flow, we have to understand, once more, that the term "*pre-judicative does not refer to something that is well formed before the 'judicative',* ***but it rather precedes the judgment****, that is, we become aware of it starting from the judgment, but we end up accepting its originality with respect to it, as if the judgment itself were fulfilled (structurally) under its condition*" (277-278), a position that is also coherent with Kahneman's "*two-system view*", but differs from it in the specific way it is applied, as "*three-system view*".

Thus, this solution model goes a step further, making "*pre-judicative memory*" "*active*" (not passive) and instrumental in the present, bridging the temporal differences, suspending the validity of the present judgement[18],

[18] "The suspension of the validity of judgments must have as its object precisely these (prejudgments active within the boundaries of the 'world'). This presupposes a veritable technique through which one can pass from logical-formal, regarding these (pre)judgments, to existential, as was stated above. How can this passing be achieved? By analyzing their

travelling back to the previous judgements through a shift from the "*formal-logic*" level to the "*existential*" level (the "*operant subject*" level) to recognize the "*pre-judgments*" by employing the "*pre-judicative memory*", identifying these bias elements and, then, returning and correcting them in the present at the conscious "*judicative*" level using the "*de-constitution operations*". The "*de-constitution operations*", at the level of the "*four-structures of the judicative form*", taking into account that the target of these operations is the realization of the "*existential leap*" that activates the "*operant subject*", will mainly engage the "*referential*" and "*existential*" structures[19]. This is the way in which "*the subject from the formal structure of the judgement*" (meaning the "*I*" from "*I am*", as in the example provided by the author) is accounted for in terms of "*reality, identity and authenticity*", these attributes ("*properties*") "*strengthening*" the "*objectual*" nature of the "*I*" from "*I am*" and thus affirming the "*personal*" dimension of the "*subject*" defined through these attributes, subject who is actually the "*operant subject*" performing these acts of judgment (280).

The movement towards "*operant subject*", by acknowledging the "*existential judgment*" (*e.g.*, "*I am*") within its incorporating "*judicative form*", fortifies its "*reality*", fact that, consequently, enhances the "*identity*" and "*authenticity*" of the subject and, at the same time, brings it into the focus of the "*evaluative judgment*" (part of the "*pre-judicative makeup*" previously discussed) applied on it, thus reaffirming the consistence and strengthening the cohesion and the "*judicative unity*" of the judgement (280). The evaluative dimension (in which our "*pre-judicative memory*" is not only a passive placeholder, but the active mechanism of exploration of past judgments that impact the current evaluations, memory that enables and conditions the "*existential*" structures (the way the memory is connected to subject's "*identity*", within the "*judicative structures*" themselves) is the one enacting the path towards "*value-driven*" judgments, in which we recuperate the "*operant*

structures and interpreting the 'leap' from the formal, from what was highlighted through analysis, to the existential, *e.g.*, to the operant subject (in the flesh, 'in person'), both operations belonging to the wider operation, also called **de**-*constitution*." (Cernica 2003, 279)

[19] "Through the operation of de-constitution of judgments, on the one hand, the 'objective' reference of the judgment, which corresponds to its expression, is subjected to a 'support' modification, being thereby ontic-ontologically 'neutralized', and on the other hand, the 'existentiality' is activated by the operant subject." (Cernica 2003, 279)

subject". This is the real actor of the judgments and reasoning pursuit, but this time enriched by the acts of experiencing and contextualizing through his/ her living existence, engaged as well in the *decision-making* processes that represent his/ her beliefs and the *goal-oriented choices and actions* that represent his/her desires, both instrumental in achieving them.

The "*act of pre-judging*" (or the "*pre-judicative act*") (Cernica 2019, 201), in this context, is using our "*pre-judicative memory*", by connecting to the past judgments of our experiences where usually the cognitive biases reside. Thus, to examine the "*object*" in scope of our present judgements, within the context of our current experiences, this act goes beyond the intentional and phenomenal into the existential aspects (199)[20], posing as well a reasoning engine that filters (same like Kahneman's "*System2*"), using the framework of past experiences, and engages the "*interpretation*" and "*re-interpretation*" of present "*non-judicative*", existential occurrences (*e.g.*, intuitive, emotional states[21]).

As we can observe, this is the "*how*" being disclosed, therefore the step further brought forward by Cernica's solution model.

The fact that any judgment or "*judicative*" construction subsumes both a "*structural*" and a "*pre-judicative*" component (Cernica 2023, 278), makes it a more effective mechanism for detecting deficiencies that might occur on either sides, to paraphrase the author (278)[22].

[20] "The interval between consciousness in the act of pre-judging and pure (or phenomenal) consciousness, viewed from a functional perspective, is the expression of an 'ontological difference': that between the phenomenon of knowledge and the very existence of a knower." And following: "At the moment this de-constitution occurs, however, through an act of consciousness (which, according to the phenomenological model, can no longer be considered intentional and constitutive phenomenal) 'someone', a 'subject', person in flesh and bones (first-person) acquires an existential profile" (201); see also 198, 201, 205.

[21] Here is Kahneman discussing the "*heuristic attribute*" of "*affective valuation*" and referencing the studies pertaining to "an automatic affective valuation – the emotional core of an attitude – is the main determinant of many judgments and behaviors." (further reference provided by the author in the original paper). The analysis continues as it follows: "In terms of the scope of responses that it governs, the natural assessment of affect should join representativeness and availability in the list of general-purpose heuristic attributes" (see Kahneman 2002, 470).

[22] "But in any of these (referring to 'structural and pre-judicative components'), deficient judicative 'forms' (structural or pre-judicative) can be found, which, in fact, present themselves as an incomplete composition of sentences (judicative 'expressions')".

While the "structural" and "pre-judicative" components unify from a "*makeup*" perspective and interact at aspectual and role levels by the mediating character of the "*pre-judicative memory*" (between "*non-judicative*" experiences and "*formal*" judgments), they continue to differ in the ways these processes and roles are applied towards the "*subject*", across its "*logical*" and "*operant*" dimensions.

For example, we consider how such processes and roles are impacting the identity of the "*logical subject*", which is established through "*formal*" "*judicative*" structures, as opposed to the identity of the "*operant subject*". The latter type of identity is defined and emerges through the subject's appeal to his own memories and experiences, being determined by all four elements of the "*judicative structure*": "*formal*", "*alethic*", "*referential*" and "*existential*".

Think how, as an example, many times the following scenario happens. You are meeting someone you haven't seen in many years, and your first encounter with them is not as the person standing in front of you now (as you ask them questions to get reacquainted) but rather as the person they appear to be in your memories, shaped by how they presented themselves all those years ago, when you last met. That is often what happens when someone tells you, after years without seeing you, that "*you haven't changed a bit*" (to paraphrase from one of Ram Dass lectures, after he commenced his journeys to India, first in 1967, and followed the enlightenment path for the rest of his life).

There is one more reason for which the "*human preferences*" would need to go often times through a suspension of present judgment, followed by operations of "*de-constitution*". These operations are using the "*pre-judicative memory*" to bring those preferences from the already assumed perceptions or intuitions domains into deliberative reconstitution, through conscious "*interpretation*" and "*reinterpretation*", taking into account the "*operant subject*" and differentiation it from the "*logical subject*" that we tend to formalize first at the "*judicative*" level. This process turns the heuristically intuitive "*human preferences*", subject to cognitive bias, into deliberative, "*value-driven*" "*human preferences*", charged, as they were recovered and reconstituted, with the attributes of "*reality*", "*identity*", "*authenticity*" and "*veracity*", as we mentioned earlier. These existential and experiential attributes are highly impactful in AI Alignment research, particularly when setting AI's "*target for alignment*", to ensure they are both adopted by AI and understood in the spirit of what is *"important"* and *"meaningful"* to transmit.

AI models, I suggest, when discussing RL (Reinforcement Learning) and IRL (Inverse Reinforcement Learning) with human feedback[23], should take into account applications of *"pre-judicative memory"* and *"pre-judicative circularity"* stances in order to be able to genuinely integrate the deliberative *"human preferences"* and expectations in their RL learning strategies.

This integration is facilitated through *"iterative"* updating of the reward functions and recurrent *"feedback loops"* across all levels outlined within the *"two-system view"* and *"three-system view"* models, employing context-dependent constrains and, as the AI models advance, *"counterfactual reasoning"*, complimented with *"adversarial learning"* (*"adversarial debiasing"*) (Yang *et al*. 2023, 55) (to detect and mitigate biases).

Such elements would serve a preventive role with regard to some of the major present concerns in AI *"Safety"* (*e.g.*, *"negative side effects"* and *"reward hacking"* risks, as instantiations of the implementation of a *"wrong objective function"*) (Amodei *et al*. 2016, 1; 3; 7; 19).

With regard to aesthetic judgments, the *"pre-judicative memory"* and *"pre-judicative circularity"* can operate with high effectiveness at the cognitive and also the experiential and existential levels, through the attributes of *"reality"*, *"identity"*, *"authenticity"* and *"veracity"* and, I would suggest, *"unicity"*. This fact is relevant, as they are, both from the creator's and observer's perspectives, subjected to a diverse historical, cultural, educational, economical contextualization that shapes and individuates both the creative processes and the aesthetic preferences. The *"judicative structures"* provide a framework for the aesthetic evaluations (*e.g.*, what is the aesthetic beauty, how is this beauty predicated, what are the temporal contexts and circumstances the aesthetic evaluations are applied to and, essentially, who is the *"operant subject"* at the aesthetic existential levels). Furthermore, they offer insight into how the subject performs based on his/her active and present *"pre-judicative memory"*, also involving the *"suspension of validity"* of certain aesthetic judgements presumed to have been infiltrated by bias. Consequently, this allows for their *"de-constitution"* and *"re-interpretation"* in order to recover their integrity, availability and credibility.

[23] Previously discussed in Sections "RL functions perspective and the link to AI Alignment" and "RL and IRL instances in AI Alignment Research", in Crețu 2025.

These process flows from Kahneman's "*two-system view*" and Cernica's "*three-system view*" are quite visibly applicable to the aesthetic judgments in the art appreciation domain. We are referring here to the plethora of "*perceptual*" and "*intuitive*" "*impressions*" that the art observer applies to the art "*object*" from both perceptual and cognitive perspectives. These "*impressions*", which often introduce bias into the process of art appreciation, also arise from the heuristics-based "*affective valuation*" as part of the "*evaluative judgments*" rooted in the "*pre-judicative makeup*", which is itself attached to the "*structural makeup*" of the whole aesthetic "*judicative*" structure. However, the above process flows are possibly even more impactful at "*first-person*" level of the artist creator domain, where the act of creation of art is tied into the "*reality*", the "*authenticity*" and the "*unicity*" of both art creative processes and art objects themselves, along with being directly linked to the "*operant subject*'s" identity. These three elements are interdependent, and one cannot function without the other. Even more so, we have to emphasize the mysterious aspects related to the aesthetic "*non-judicative*" ("*non-cognitive*") domain, where the affective charges are pushed to the limit, and in which the experiencing is prior to judgment, feeding into a multitude of intuitions that flood, through the representational content layers, the "*judicative*" domain. Along with the "*judicative*" regularisation, the "*non-judicative*" influences gain, as well, a lot of traction both at the art observer's level and in art appreciation processes.

At the art observer's level, the aesthetic "*preferences*" would need to undertake the same purifying process of "*de-constitution*", "*interpretation*" and "*re-interpretation*", to be able to transition from "*intuitive preferences*" towards more educated, "*deliberative preferences*".

With regard to the processes of art appreciation, the "*reward*" is connected to the "*pleasure*" of seeing and sharing the art experience. However, as we evolve towards more elevated "*cumulative*" aesthetic "*rewards*", progressing through successive phases of "*de-constitution*" and "*interpretation*" via iterative "*pre-judicative circularity*", the "*pleasure*" begins to depart from a "*reward*" incentive-based attitude and transforms towards "*value*". This is the moment when aesthetic "*value*" becomes embedded in the experience of art appreciation. This may be a lengthy process, but its payoff could be more significant, yielding a more fulfilling art experience for the art observer from a reflective standpoint (being able to appreciate the insights in a work of art and to connect to the artist's imagination and creative experience), a cognitive standpoint (being able to grasp the conceptual elements showcased throughout art representations, to understand their contents and, also, to effectively evaluate them, by comparison with others) and an interpersonal standpoint (being able, as an art observer participating in the processes of art appreciation, to be part of the broader art community and become an art lover, supporting creative endeavors and the betterment of society through art).

For artist creator, the dynamics of "*pleasure*" and "*reward*" and many times, "*pleasure*" versus "*reward*", are transformative of the art experience he/she undergoes. From this perspective, the "*atypical salience attribution*"[24] permeates the artist's endeavor and impacts directly the artist's aesthetic processes and judgments. Thus, the pursuit of art is driven by "*rewards*" (*e.g.*, "*wanting*" to create a work of art or a piece of music) that progressively leave the territory of pleasure (*e.g.*, "*liking*" to paint or to write music), at least in its hedonic sense, towards one of discovering and capturing a different kind of "*significance*" or "*meaning*" of what is witnessed in the world (this one or other worlds that may be, empowered by the forces of imagination) and then taken from it (or from them) into what is being manifested and expressed through the art object. With this different frame of reference, the aesthetic judgments of the artist creator shift from a "*reward-based*" model, in which reward is equated with pleasure, towards a "*significance- and meaning-based*" model, wherein he/ she approaches a "*value-driven*" experience of art. Such an experience unfolds both at the "*first-person*", intrapersonal, metacognitive level (*the wanting it* from "*reward*" turns into *the being it* from "*self-actualization*", to recall Maslow's term) and at the "*third person*", interpersonal, intersubjective level (*the wanting it* from "*reward*" turns into *sharing* it and *communicating* it within the "*shared manifold of intersubjectivity*", to recall Gallese's hypothesis).

By this time, the hedonic pleasure from the nascent phases of art creation (*e.g.*, "*I am painting because it relaxes me*") has transformed into "*value-driven*" and "*goal-oriented*" fulfilment and the creation of art has become an existential practice. It is the practice in which the "*non-judicative*"(instead of being a remote place at the fringes of our sight,

[24] See Crețu 2024; in what concerns the relationship between "*liking*" vs. "*wanting*"; see also Gross 2024, 1.

hidden in the heuristic realm, where biases lie in wait, ready to surface into the cognitive space) becomes, through a conscious "*de-constitution*", "*interpretation*" and "*re-interpretation*" effort, a powerful resource that the artist creator starts mastering. Thus, the "*non-judicative*" domain, while maintaining its specificity, moves closer to and integrates more fluidly with the "*judicative*" domain. Particularly for creative processes, "*abductive reasoning*" and "*analogical reasoning*" are employed in order to integrate "*symbolic knowledge*", while "*intuitive reasoning*" is used in order to integrate "*episodic knowledge*". Both "*symbolic knowledge*" and "*episodic knowledge*" types can intersect at "*intrapersonal reasoning*" level[25], the latter bridging the gap between them, that for others, perhaps less oriented towards creative endeavors, tends to remain an unpassable chasm.

This synergy creates a powerful rippling effect. It contributes to the artist creator becoming aware of dimensions otherwise inaccessible to the cognitive mind, while also making the artist creator aware of his/ her own inner completeness. This, in turn, empowers this artist, enhances the creative processes and allows him/her to be in the flow, in the act of artistic creation. Furthermore, this synergy advances the artist creator along the "*value-driven*" path, enabling him/her to aim towards heightened states of awareness and more evolved and refined solutions and goals.

References

American Psychological Association (2025). *APA Dictionary of Psychology*, available online at <https://dictionary.apa.org/> last time accessed on March 29, 2025.

Amodei, Dario; Chris Olah; Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. (2016). "Concrete Problems in AI Safety". *arXiv*, available online at < http://arxiv.org/abs/1606.06565 >, last time accessed on July 25, 2016.

Cernica, Viorel (2013). *Judecată și Timp. Fenomenologia Judicativului*. Institutul European.

Cernica, Viorel (2019). „Actul de a pre-judeca şi re-constituirea subiectului. Aplicație în hermeneutica pre-judicativă." In Viorel Cernica (ed.), *Studii în hermeneutica pre-judicativă și meontologie* (*Studies in Pre-Judicative Hermeneutics and Meontology*), Vol. 4. Bucharest: University Press, 185-249.

[25] In what concerns the analysis of various types of knowledge see the Ben Goertzel's AGI (Artificial General Intelligence) model, presented in my paper, "The Advent of AGI" (Crețu 2023, 124-147).

Cernica, Viorel (2023). „Despre o aporie identitară: Eu sunt non-ființă" ("About an Identity Aporia: I am non-Being"). In Viorel Cernica (ed.), *Studii în hermeneutica pre-judicativă și meontologie* (*Studies in Pre-judicative Hermeneutics and Meontology)*, Vol. 8. Bucharest: University Press, 267-298.

Cretu, Sorin (2023a). "Heuristics, cognitive biases, metaphors, embodied mind. The road to being human." In *Studii în hermeneutica pre-judicativă și meontologie (en. Studies in Pre-Judicative Hermeneutics and Meontology).* Vol. 7, coordinated by Viorel Cernica. Bucharest: University Press.

Crețu, Sorin (2023b). "The Advent of AGI." In *Studii în hermeneutica pre-judicativă și meontologie (en. Studies in Pre-Judicative Hermeneutics and Meontology).* Vol. 8, coordinated by Viorel Cernica. Bucharest: University Press.

Crețu, Sorin (2025, forthcoming). "Into a new era of 'reward-punishment' or towards the liberation from it?" In *Studii în hermeneutica pre-judicativă și meontologie (en. Studies in Pre-Judicative Hermeneutics and Meontology).* Vol. 9, coordinated by Viorel Cernica. Bucharest: EIKON Publishing.

Gallese, Vittorio (2003). "The Manifold Nature of Interpersonal Relations: The Quest for a Common Mechanism." In *Philosophical Transactions: Biological Sciences* 358 (1431):517-28.

Gross, Madeleine E., James C. Elliott, and Jonathan W. Schooler (2024). "Why Creatives Don't Find the Oddball Odd: Neural and Psychological Evidence for Atypical Salience Processing." In *Brain and Cognition* 178 (August 2024): 106178. https://doi.org/10.1016/j.bandc.2024.106178.

Kahneman, Daniel (2002). "Maps Of Bounded Rationality: A Perspective on Intuitive Judgment and Choice," *Prize Lecture*, December 8, 2002, Princeton University, Department of Psychology.

Yang, Jenny, Andrew A. S. Soltan, David W. Eyre, Yang Yang, and David A. Clifton (2023). "An Adversarial Training Framework for Mitigating Algorithmic Biases in Clinical Machine Learning". In *Npj Digital Medicine* 6 (1):55. https://doi.org/10.1038/s41746-023-00805-y.

All links were verified by the editors and found to be functioning before the publication of this text in 2024.

DECLARATION OF CONFLICTING INTERESTS

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

FUNDING

The author received no financial support for the research, authorship, and/or publication of this article.

Creative Commons Attribution 4.0 International License

https://annalsphilosophy-ub.org/2024/03/2-copyright-statement/