

# Epistemic considerations when AI answers questions for us

Johan F. Hoorn<sup>1,2</sup> and Juliet J.-Y. Chen<sup>2,\*</sup>

<sup>1</sup> The Hong Kong Polytechnic University, School of Design and Dept. of Computing

<sup>2</sup> Laboratory for Artificial Intelligence in Design (AiDLab)

\* Corresponding author: jychen [at] aidlab [dot] hk

## Abstract

In this position paper, we argue that careless reliance on AI to answer our questions and to judge our output is a violation of Grice's Maxim of Quality as well as a violation of Lemoine's legal Maxim of Innocence, performing an (unwarranted) authority fallacy, and while lacking assessment signals, committing Type II errors that result from *fallacies of the inverse*. What is missing in the focus on output and results of AI-generated and AI-evaluated content is, apart from paying proper tribute, the demand to follow a person's thought process (or a machine's decision processes). In deliberately avoiding Neural Networks that cannot explain how they come to their conclusions, we introduce logic-symbolic inference to handle any possible epistemics any human or artificial information processor may have. Our system can deal with various belief systems and shows how decisions may differ for what is true, false, realistic, unrealistic, literal, or anomalous. As is, stota AI such as ChatGPT is a sorcerer's apprentice.

## 1. Why would you?

"Methods and tools are for those who are fools." Another saying would be: "If you can't make it, fake it." The creation of Deepfakes as well as systems like Dall-e and the GPT series make it all too easy to be lazy and carelessly rely on Neural Networks to produce content. At first sight, such output seems pretty harmless if produced for art or entertainment purposes. Still, we humans are more vulnerable than we think. Even though we subconsciously 'know' what is fake, our brains mitigate the premise of fake and subconsciously naturalize the content of the information it carries as part of what 'I' perceive to be 'mine.' Memory, which is almost impossible to erase with specificity, can be subliminally edited.

The charm of a Neural Network is that to the unsuspecting eye, what is generated seems to be relevant and coherent. Yet, we can trace back and identify its causes with little degree of certainty although a mass exchange of information is going on from person to person, media to person, media to media, all at the same time. Countless pieces of information are intertwined, but this is so hard to stomach that we give up looking for a comparatively convincing 'version of the truth.'

As new information is incorporated into memory, the contents of memory are increased, and memories enriched and modified. People are often unaware of the impact that information intrusion can have on individual and collective memory (Loftus, 1993). This may lead to existential and identity crises, in which people fall into a self-perceived emptiness, where their feet cannot reach the ground. When the fake is cheaper than the real, then the real become so thin. After all, what we perceive and call 'Reality' (Section 7) also consists of conditional untruths. On second thought, then, careless usage of Neural Networks becomes harmful when output is

claimed as factual descriptions of real life situations, which immediately begs the question what factual and real may be at all.

Why people do this is to gain a competitive edge by getting better grades, improving their CVs, achieving results they otherwise would not be capable of. *Per capita*, then, AI will improve everybody's output so that eventually, we are all equal again. The point is, however, that by using AI in this manner, we wear high heels so we can reach higher but we did not grow taller. By that, we create a world of pretense that becomes increasingly harder to navigate from an epistemic point of view: What the mediated messages convey points to nothingness.

We are creating a new filter bubble. Large Language Models produce an awkward world view from biased training sets muddled up with hallucinations because the error term cannot be reduced; people put those concoctions as self-produced texts on the net, which are then picked up by Large Language Models to confirm the awkward world views created from a jumble of even heavier biased training sets and unidentified errors.

'When I see the generated result, then that is what I really thought.' ChatGPT is influential for the moral judgement of users, even if it does not provide coherent advice. Despite knowing that they are communicating with a chatbox, users underestimate the extent to which their judgement and decisions are influenced. People's confidence in being true to themselves remains high, but their moral judgement and ability to make independent decisions becomes impaired (Krügel, Ostermaier, & Uhl, 2023).

As a further consequence, those who openly rely on systems like ChatGPT will lose authority. They will lose authority because either the Neural Network is too hallucinatory, biased, or silent about its sources, *or* because the system becomes so good that it outperforms its user. Either way, the user loses credibility and authority. Nowadays, Microsoft mixes advertisements in its 'new Bing' so one will get scientific summaries with stealth advertising blended in from who pays the most. So why would we? We are heading for a communication crisis rooted in default disbelief of mediated content because it could be AI-generated and thus, default unreliable.

## 2. The communication maxim

Among the maxims of communication that Grice (1989) distinguishes, GPT systems answering our questions may be said to fall short on all four, but the Maxim of Quality is the most important to our deliberation:

Quantity	Provide the right amount of information (no more no less)
Quality	Give an honest rendition of your beliefs (tell the truth)
Relation	Information should be relevant to the goals of a conversation
Manner	Be clear to your audience (no jargon, no beating about the bush)

One of the conversational rules following from the Maxim of Quality is that one gives the speaker ample credit to speaking the truth and suspend disbelief until something counter-factual occurs and the Maxim of Quality is violated (from that listener's perspective). With AI that cannot be fact-checked, we are heading for a reverse conversational rule to the Maxim of Quality: Default is fake. The AI or the person using AI should provide evidence to each statement so that the conversation partner can believe the artificial interlocutor. That, probably, should be non-mediated evidence, which means that job interviews and exams are going to be face-to-face and

without any digital means allowed just to hear what insights the candidate actually has. Personal insights, not just knowledge that can be memorized or looked up.

Going digital has become a big cheat box. That in AI communication central concerns of honesty as opposed to deception are at stake brings back reminiscences of social-media theories such as Signaling Theory (Donath, 2007) and Warranting (e.g., Walther, Van der Heide, Hamel, & Shulman, 2009). AI responses to human questions are *conventional* signals (Donath, 2007); they are abstract and symbolic (i.e. digital words and images) and so they are less reliable than *assessment* signals (ibid.), which are directly accessible. *Conventional* signals need more fact-checking and warranting because they cue that, for instance, the assignment turned in online is from the student, which is different from actually seeing the student write that assignment with pen-and-paper in front of the teacher (*assessment* signals).

### 3. Frequentists administering the Turing test

In the AI community, the standard approach to decide for human or machine-generated work would be to run a Turing test, and according to the master himself (Turing, 1950), if the ‘imitation game’ turns out indistinguishable results, the machine may be considered perceptually ‘intelligent,’ ‘conscious,’ or show any other human quality of interest. However, Turing merely speaks from a phenomenological standpoint that the machine is *psychologically* perceived as such. Searle (1980) would argue that the information processor in the Chinese Room neither is ‘intelligent’ nor ‘conscious,’ despite impeccable output. Unfortunately, the imitation game rests on a *fallacy of the inverse* or *inverse error* ( $A \rightarrow B, \neg A, \rightarrow \neg B$ ), denying the antecedent while inferring the inverse from its original statement (Hoorn & Tuinhof, 2021):

If I see differences with humans (A), the agency is artificial (B)  
I do not detect differences with humans ( $\neg A$ ) (missing the signal)  
Then the agency is not artificial ( $\neg B$ )

However, there is more at hand. The performance of an AI may not be distinguishable from that of a human but is it the same? Did the student write from her own insights and learning or did her Large Language Model just summarize other people’s texts (not thoughts)? In standard approaches to test theory, Student *t*-tests or derivations thereof, would decide for difference beyond doubt (or not) given a pre-set rejection area. With the student using very good AI, frequentist approaches would fail to reject the  $H_0$  and with Turing (1950) conclude there is no detectable difference between the probability for AI-generated work and the probability for human-made. However, they are still two distributions not drawn from the same population, Searle (1980) would argue, but the test just was not sensitive enough to show the difference beyond reasonable doubt. It would be a fallacy, then, to conclude for sameness and think the student did her work well.

### 4. The legal maxim

Everything is turned upside down when AI does the fact checking for us. If the student turns in her own work, but the AI-enhanced plagiarism scanner deems the similarity rate with AI-

generated copycat work too high, then all of a sudden, the student has the burden of proof that s/he did not cheat. In absence of opposing evidence, the student will be found guilty.

That breaches one of the fundamentals of Western legal systems, dating back to 13<sup>th</sup> century Cardinal Jean Lemoine, formulating that *item quilbet presumitur innocens nisi probetur nocens*: a person is presumed innocent until proven guilty (Ullmann, 1950; Pennington, 2003), after which the Inquisition would torture people to a coerced confession – *ad baculum*. Coined by Locke, *argumentum ad ignorantiam* is an appeal to ignorance by which is meant the absence of counter-evidence. Although structurally the formal logics may be correct and although empirically the criminal may be guilty (Walton, 1992), in informal logic, an appeal to ignorance is a fallacy but a justified one because the prosecutor who makes the claim also has the burden of proof, else it would not be ‘honest’ or ‘fair,’ leading to everyone accusing everybody else, leading to great social unrest. Take that as a warning for believing current plagiarism scanners because they do not handle their *t*-tests right (if at all), neither according to Fisher (1932), Neyman-Pearson (1933), Lindquist (1940), nor to Bayes (Jeffreys, 1961, p. 283).

The legal inference of an appeal to ignorance is that what has not been proven false must be true. In the case of the plagiarism scan (or scam?), the prosecutor does so to the defendant’s disadvantage: The student is guilty until proven not so. That is a simplistic bipolar conception of just one type of epistemics: true = 1 – false. In a unipolar conception, true and false can co-exist and affect each other but not necessary sum up to 1; they interact in dynamically varying ratios (cf. an audio equalizer). Yet another epistemics would say that truth and untruth make a bi-dimensional unipolar scale that is related to false and not-false, another bi-dimensional unipolar scale, together lingering in a kind of superposition, being ambiguous about the final verdict while the juridical process is still underway (Ho & Hoorn, 2022).

## 5. The reverse Turing test

In reversing the Turing test, the AI decides whether it interacts with another computer or a human being (Feigenbaum, 2003); whether the presented work is AI-generated or human-made (or mixes thereof). Although persuasive maybe, we move the syllogism of *argumentum ab auctoritate* (Fellmeth & Horwitz, 2009) from the human to the ‘expert’ AI: We draw the argument from authority from a machine.

The concept of expertise proofs crucial. Although fallacious nonetheless, if the authority is the source who holds, to the best of our knowledge, the most expertise, then an appeal to authority may be regarded as non-fallacious. Maybe. Because under whatever epistemics or epistemology that is in operation, relying on authority remains a matter of belief, induction, probability, and likelihood. Proof can only be delivered in closed-state systems such as classic logics and mathematics, not through anything that relates to open systems or that is vulnerable to information entropy. Full fallacy is committed when the belief is in a non-authoritative source, one that garners but little consensus, such as the GPT series that according to their developers:

... is incredibly limited but good enough at some things to create a misleading impression of greatness. It’s a mistake to be relying on it for anything important right now. It’s a preview of progress; we have lots of work to do on robustness and truthfulness. (Sam Altman (@sama), CEO of OpenAI, Dec. 11, 2022, <https://twitter.com/sama/status/1601731295792414720?lang=en>)

Reversing the Turing test, then, as in the case of the plagiarism scanner, is a full authority fallacy and moreover, a violation of the legal maxim of innocence if the burden of proof is not put on the system, which should not be telling crude similarity but telling who is copying who. Additionally, a frequentist approach to the reversed Turing test commits a Type II error (i.e. missing the signal) if the plagiarism scanner concludes for sameness, resulting from the *fallacy of the inverse* that not detecting difference with AI-generated work would mean the work is not the student's own (Hoorn & Tuinhof, 2021):

Your own work would be different from AI  
Not different?  
Then not your own work

The student is guilty until proven not so.

## 6. What is missing in all of this

What is missing in all of this mess of people misusing AI is proper citation. People can sample as much music as they like as long as they say where they got it from (and in protectionist societies, pay the bill). That is good open-source practice, conducive to creativity and innovation, building a continuous body of scientific work where we can see who is standing on “ye shoulders of Giants” (Newton, Feb. 5, 1676, in a condescending letter to Robert Hooke).

Yet, how small a chunk of sampling is still tolerable? If one breaks a piece of music down to its singular notes, the composer still copies from that piece of music although not using the same constellations. The same goes for words. To explain a theory, science demands that one technical term has one and but one meaning alone. The same goes for the use of equations and formulae. Haphazardly using synonyms or paraphrasing equations so to avoid plagiarism is merely introducing confusion if not downright academic misconduct.

Furthermore, against all Anglo-Saxon management practice of today, ‘factuality’ and ‘truthfulness’ are not about the end product, result, revenue, or output but about the thinking process, the throughput. Indeed, the journey *is* the destination (Lao Tzu in *Tao Te Ching*). Die Wanderlust, not the goal director. Die Wunderlust, not the target setter. This creed may seem too romantic and philosophical to some but has hard-core consequences, those the rational economists of science, looking to maximize their academic profit, may not like: Without realizing it, they are caught in an epistemics of the virtual.

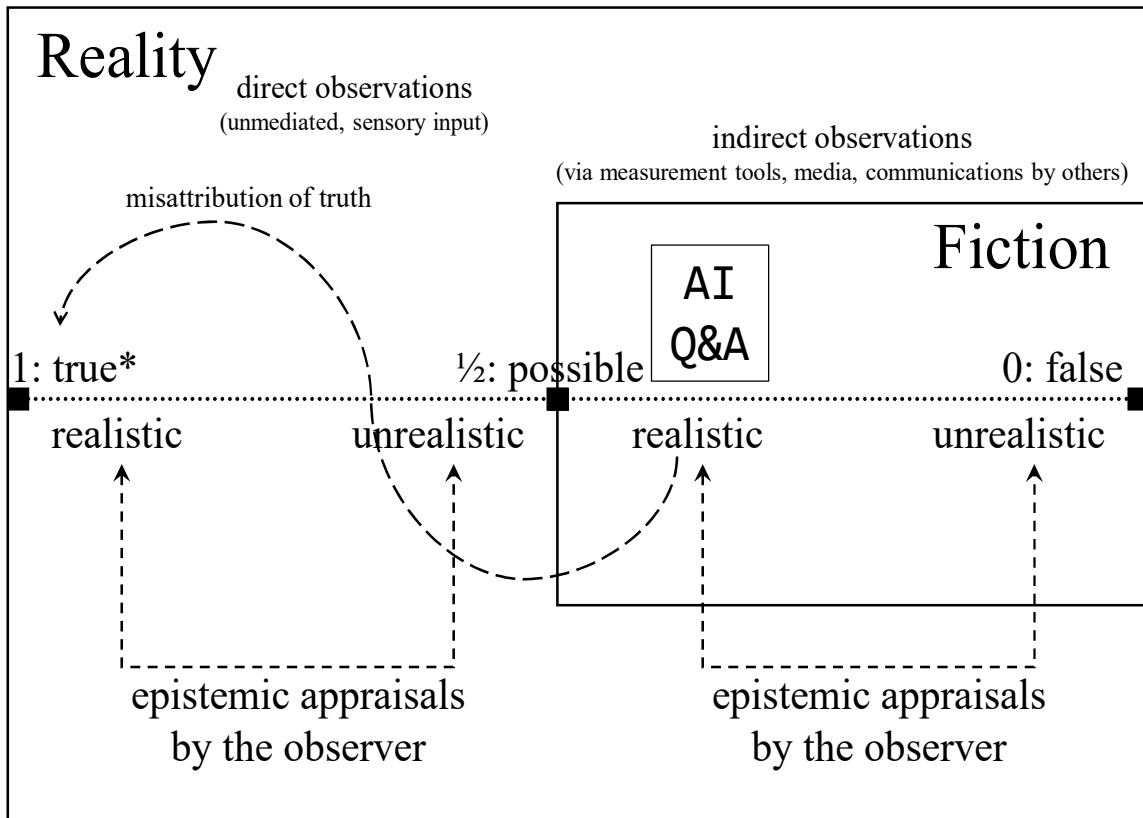
## 7. An epistemics of the virtual

In the Laboratory for Artificial Intelligence in Design (Hong Kong SAR), we are developing *EpiVir*, which is shorthand for and an implementation of Epistemics of the Virtual (Hoorn, 2014), a logic-symbolic inference system that keeps ontological classifications in check with elaborate epistemic appraisal processes. Figure 1 shows the conceptual layout of the system, while the logic-symbolic structure can be found in Hoorn (2014, Chap. 7).

Figure 1 supposes that the physical world can be assessed by sensors but in a limited and biased way. Therefore, what is seen as Reality is conceptual and changeable and part of that Reality is explicitly marked as Fiction because mediated content such as novels and feature films

exist in that conception of Reality. However, in Reality, statements are assumed to refer to things that are present in the physical world in the here and now whereas in Fiction, they are empty references: Things that do not exist, partially exist, have existed but not anymore and so are out of observational reach, or things that may come into existence but no one knows for sure.

## Physical Universe



\*Attribution of truth according to observer's belief system or world view (e.g., scientific, religious, or cultural)

Figure 1. Conceptual framework of Epistemics of the Virtual (*EpiVir*).

Yet, Fiction may be very realistic. Although as of yet, no one directly observed black holes, dark matter, white matter, or a white dwarf, not the least of scientists strongly believe that a wormhole can be created on a quantum chip to teleport a subatomic particle (Jafferis et al., 2022) whereas others shame those same scientists for having created a hoax. In the same vein will non-experts believe the highly realistic renderings of the GPT series, which are mediated summaries of mediated messages by unknown sources and hence none but realistic Fiction. Non-experts on the topic will misattribute truth to those industrialized AI mass-productions of content, notwithstanding (Figure 1, AI Q&A).

As is, *EpiVir* is capable of maintaining a database with previously stored knowledge for which parameters can be set to allow changes to that knowledge base as a function of deviation tolerance, suspension of disbelief, and a variety of decision lemmas (e.g., classic or fuzzy logics, quantum probability). The settings may vary, the weights of the features that are processed may change, but the system does not. That is the claim: Epistemics of the virtual is the path all organic

and artificial information processors have to walk to assess the world about us, whether they are profit-maximizing rationalists, journey-loving freethinkers, or ChatGPT, whether they are faulty error-makers or highly accurate administrators. For one information processor, *EpiVir* merely runs ontological classification, making it error-prone when hybrids and exceptions need to be handled (think of neural net classification mistakes); for another information processor, *EpiVir* allows many deviations to the classification templates and conceives of the world as mere hypotheses and possibilities, taking hardly anything for ‘real.’

In other words, *EpiVir* can work with any sort of reality perception because that is merely the input to the throughput, depending on subjective belief systems or world views (e.g., scientific, religious, educational, or cultural) on the basis of which ‘truth’ is attributed to an observed entity (Figure 1). Thus, Earth is flat, spherical, or an irregular ellipsoid are three different ontologies that *EpiVir* can handle with the same ease, representing different world views or perspectives. It can adapt from ‘flat’ to ‘irregular ellipsoid’ by widening criteria of acceptance and tolerance to deviation (or vice versa, narrowing down what is acceptable).

Not only will *EpiVir* label an observation as true, false, or ‘probable’ (according to belief), for things it is not sure about, it may attribute a level of being realistic or not, or even, when the settings allow so, reckoning certain statements about the world as metaphorical (e.g., Earth is a blue marble). A  $\beta$ -version in Python of the *EpiVir* API is available on the GitHub and the `main.py` and `signal_detection.py` components can be found in Appendix 1.

## References

- Donath, J. (2007). Signals in social supernets. *Journal of Computer-Mediated Communication*, 13(1), 231-251. doi: 10.1111/j.1083-6101.2007.00394.x
- Feigenbaum, E. A. (2003). Some challenges and grand challenges for computational intelligence. *Journal of the Association for Computing Machinery*, 50, 32-40.
- Fellmeth, A. X., & Horwitz, M. (2009). *Guide to Latin in international law*. Oxford, UK: Oxford University.
- Fisher, R. A. (1932). Inverse probability and the use of likelihood. *Mathematical Proceedings of the Cambridge Philosophical Society*, 28(3), 257-261.
- Grice, H. P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University.
- Ho, J. K. W., & Hoorn, J. F. (2022). Quantum affective processes for multidimensional decision-making. *Nature: Scientific Reports*, 12, 20468. doi: 10.1038/s41598-022-22855-0
- Hoorn, J. F. (2012). *Epistemics of the virtual*. Philadelphia, PA: John Benjamins.
- Hoorn, J. F., & Tuinhof, D. J. (2022). A robot’s sense-making of fallacies and rhetorical tropes. Creating ontologies of what humans try to say. *Cognitive Systems Research*, 72, 116-130. doi: 10.1016/j.cogsys.2021.10.003
- Jafferis, D., Zlokapa, A., Lykken, J. D., Kolchmeyer, D. K., Davis, S. I., Lauk, N., ... & Spiropulu, M. (2022). Traversable wormhole dynamics on a quantum processor. *Nature*, 612(7938), 51-55.
- Jeffreys, H. (1961/1939). *Theory of probability*. Oxford, UK: Clarendon.
- Krügel, S., Ostermaier, A., & Uhl, M. (2023). ChatGPT’s inconsistent moral advice influences users’ judgment. *Nature: Scientific Reports*, 13(1), 4569.
- Lindquist, E. F. (1940). *Statistical analysis in educational research*. Boston, MA: Houghton Mifflin.

- Loftus, E. F. (1993). The reality of repressed memories. *The American Psychologist*, 48(5), 518-537.
- Neyman, J., & Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706), 289-337.
- Pennington, K. (2003). Innocent until proven guilty: The origins of a legal maxim. *Jurist*, 63, 106.
- Searle, J. R. (1980) Minds, brains, and programs. *Behavioral Brain Science*, 3(3), 417-457.
- Turing, A. M. (1950). I.—Computing machinery and intelligence. *Mind*, LIX(236), 433-460. doi: 10.1093/mind/LIX.236.433
- Ullmann, W. (1950). The defense of the accused in the medieval inquisition. *The Irish Ecclesiastical Record*, 73, 481-489.
- Walther, J. B., Van der Heide, B., Hamel, L., Shulman, H. (2009). Self-generated versus other-generated statements and impressions in computer-mediated communication: A test of warranting theory using Facebook. *Communication Research*, 36, 229-253.
- Walton, D. (1992). Nonfallacious arguments from ignorance. *American Philosophical Quarterly*. 29(4), 381-387.

## Acknowledgements

This study is funded by the Laboratory for Artificial Intelligence in Design (Project R2P3), Innovation and Technology Fund, Hong Kong Special Administrative Region. We are grateful to Johnny K. W. Ho and Yuwei Tang for their implementation of the *EpiVir* code.

## Appendix 1

*EpiVir* software ([main.py](#) and [signal\\_detection.py](#)). Available from <https://github.com/letokanoce/EOTVRESTAPI.git>

[main.py](#)

```
# ontological classification

import numpy as np
import pandas as pd

from epbelsys.entity import PerceivedEntity, EntityProfile
from epbelsys.model import EnvironSettings

class BaseProfile:
    def __init__(self, feature_set, p_value, weight, corr):
        self.feature_set = np.array(feature_set)
        self.p_value = np.array(p_value)
        self.weight = np.array(weight)
        self.corr = np.array(corr)

    def show_profile(self):
        data = {
            'Feature': self.feature_set,
            'P-Value': self.p_value,
            'Confidence': self.weight,
            'Correlation': self.corr
```



```

    }
    return pd.DataFrame(data)

class CacheMediator:
    def manipulate_cache(self):
        pass

class RedisManipulation(CacheMediator):

    def manipulate_cache(self):
        # implementation to manipulate cache using Redis
        pass

class MatchMediator:
    def manipulate_match(self, base_profile, entity_profile):
        pass

class SDTManipulation(MatchMediator):
    def manipulate_match(self, base_profile, entity_profile):
        # implementation to manipulate match using SDT and entity_profile
        pass

class OntologicalClassificaton():
    def __init__(self, perceived_entity: PerceivedEntity, settings: EnvironSettings):
        self.entity_profile = self._construct_entity_profile(perceived_entity)
        self.cache_mediator = CacheMediator()
        self.match_mediator = MatchMediator()
        self.settings = settings

    def _construct_entity_profile(self, perceived_entity):
        feature_set = perceived_entity.features
        p_value = perceived_entity.p_value
        confidence = perceived_entity.p_val_con
        return EntityProfile(feature_set, p_value, confidence)

if __name__ == "__main__":
    settings_1 = EnvironSettings("reality", "literal")
    features = ["noes", "eyes", "tail"]
    p_value = [0.62, 0.3, 0.01]
    pval_con = [0.9, 0.5, 0.6]
    perceived_entity_1 = PerceivedEntity(features, p_value, pval_con)
    oc_1 = OntologicalClassificaton(perceived_entity_1, settings_1)
    print(oc_1.cache_mediator)

```

### signal\_detection.py

```

import numpy as np
from scipy import stats

from epbelsys.model import BaseProfile

class SignalDetection:
    def __init__(self, profile: BaseProfile):
        self.profile = profile
        self.z_hit = -stats.norm.ppf(self.profile.p_hit)
        self.z_fa = -stats.norm.ppf(self.profile.p_fa)
        self.mean_z_hit = np.mean(self.z_hit)
        self.mean_z_fa = np.mean(self.z_fa)
        self.beta_tp = 0.8

```

```

@property
def sigma(self):
    return stats.linregress(self.z_fa, self.z_hit).slope

@property
def d_prime(self):
    return np.sqrt(2 / (1 + np.square(self.sigma))) * (self.mean_z_hit + self.sigma * self.mean_z_fa)

@property
def criterion(self):
    return np.sqrt(2 / (1 + np.square(self.sigma))) * (self.sigma / (1 + self.sigma)) * (
        self.mean_z_hit + self.mean_z_fa)

def cal_c_tp(self):
    if self.sigma != 1.0:
        return (-self.d_prime + self.sigma * (
            np.sqrt(np.square(self.d_prime) + 2 * (np.square(self.sigma) - 1) *
np.log(self.beta_tp)))) / (
            np.square(self.sigma) - 1)
    else:
        return (np.square(self.d_prime) / 2 + np.log(self.beta_tp)) / self.d_prime

```