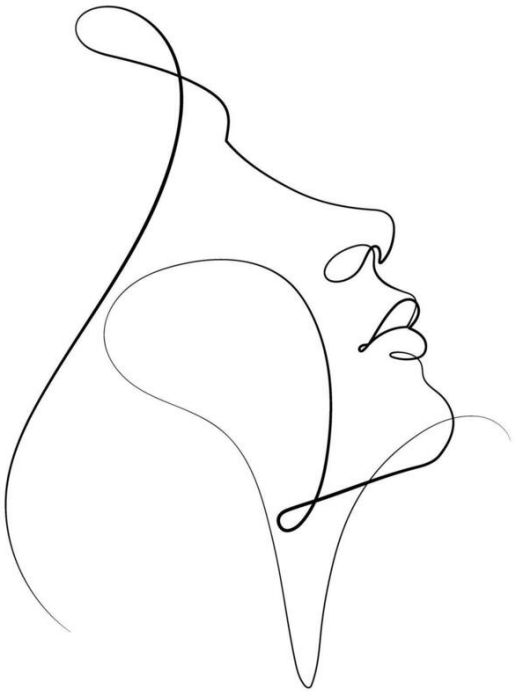
The Human



Lola Huang, Maria Kostylew

Contents

1 Preface

I met Maria Kostylew in the summer of 2025 at the Summer Program on Applied Rationality and Cognition (SPARC). She was an instructor; I was a first-year participant. Although we had only a three-hour one-on-one conversation, I admired her approach to knowledge and her deep desire to steer societal progress in the age of AI. After SPARC, we continued our discussions on a 2-person server, exploring our shared ethical standards, ways of knowing, and general goals, while debating our minor and major differences. Our talks often left me with the need to journal – to scrutinize my own knowledge structures and let the conclusions of our arguments reshape my thinking. This compilation presents some of our conversations and my subsequent reflections in a structured, dialectic form, inspired by Plato's *The Republic*. Unlike Plato, I try to enter each discussion with as few presumptions as possible, acting not as a teacher, but as an interviewer (or more precisely, an interrogator.)

Boldly titled *The Human* without Maria's consent, this collection spans topics first sparked by fragments of our in-person conversation:

(i) Thinking about thinking;

(i) Music;

(i) Mathematics;

(i) Alignments between goals of life and actions at present;

(i) Epistemology and Metaphysics.

We are aware that our conversations may eventually reach beyond anthropocentric concerns. Yet I retain the title *The Human*, because this is not a study *of* humans, but an introspection into humanity – humanity's interaction with its surroundings, and its future paths, seen through the lens of two individuals. We do not claim to represent the human condition in full, but we believe in the power of this document to spark curiosity. We hope to inspire others to examine their own beliefs, ethics, and worldviews with similar fervor, and to popularize an uncommon form of intellectual friendship.

We present the discussions in the chronological order the questions first arose. When multiple questions branch from one response, we order them by when they were addressed. For questions or comments, please email huangxinyan_lola@outlook.com.

The document was last updated on Saturday 25th October, 2025.

Lola Huang

2 Thinking about Thinking – Part I.

"Poirot," I said. "I have been thinking."
"An admirable exercise my friend. Continue it."

Agatha Christie, Peril at End House

2.1 Opening Question

The mind is of or about things. Many contend that intentionality is one of the two primitive features of the mind. Notably, however, we restrict ourselves to some tangible intentionality when we think about *the things we think about*. To illustrate this claim, we encourage the readers to engage in the following exercise:

(i) Think.

(ii) Write down, in words, what you have been thinking about. Make sure you have included, to a maximal extent, the thought processes, sensations, and [possibly inexplicable] sparks that has occurred through your act of thinking.

(iii) Answer the following question: give a mutually-exclusive and collectively-exhaustive list of *things* that the mind could be of or about.

(iv) Beyond material objects and first-order abstract entities [e.g., emotions, theories of concrete things,] what else does your list consist of?

(v) Does your list embody a form of meta-level thinking, coined as "thinking about thinking" by us? If so, what are instances of thinking about thinking for you?

We shall highlight the value of thinking in the meta as hinted at by the fifth exercise. When one thinks about thinking, she questions her underlying system of beliefs much like the actions of a plumber: the latter installs and maintains often-neglected physical systems, while the former scrutinizes her mental structures which has developed through a combination of nature, nurture, and that of chance. Thinking about thinking also allows one to see the bigger picture before getting lost in the immediate details. In fact, the following dialogues serve as an effective exercise for the author themselves – they allow us to make explicit implicit assumptions, thought processes, and opinions we have otherwise taken for granted. Last, the value of thinking about thinking permeates the post-thinking phase, as it allows alignments and regulations where one adjusts her thought processes based on reflection, inducing strategy-switching and goal-reorientation.

Lola: Does "thinking about thinking" occur simultaneously for you? How does reflection usually emerge? Personally,

(i) If I am dealing with a familiar topic, then my zeroth-order reflection (plain thinking) and first-order reflection (thinking about thinking) occur spontaneously.

(ii) For unfamiliar topics, I find myself circling around the subject matter for quite a few times without first-order reflection, yet each reflection gives rise to new insights that ultimately allow first-order reflections to emerge. For me, reflection is a strongly-emergent property.

Maria: For me there are a couple of main types of thinking, and first-order reflections occur for each of them. *Maria does not address the "thinking about unfamiliar topics" case explicitly, but I think we could attribute (i.) to thinking about unfamiliar things.*

(i) Whilst I am speaking, like while actively trying to figure things out, there is an internal monologue — "the kind of deliberateness where words feel more tactile as I arrange them."

(ii) If I am familiar with something, words come out spontaneously and intuitively, but the "thinking about thinking" loop is almost always there (irrelevant to my familiarity on the topic) – as long as I am consciously thinking about something. I notice it in between breaths in the midst of thoughts, much like glimpses of an eagle eye view.

(iii) Sometimes my thoughts look more like "drifting through different colorful clusters of random concepts and sensory stimuli." In this case, thinking about thinking happens, also in glimpses, but feels less salient and more leisurely.

I should highlight some post-discussion reflections that emerged while rewriting the messages.

First, we have never defined thinking about thinking (TaT). Here's my definition:

Definition 2.1. Thinking about thinking refers to the conscious act of reflecting upon thoughts on a specified subject.

We could expand upon Maria's association of a bird-eye view to the conscious act of thinking about thinking. Let's metaphorically refer to the entity that thinks about thinking as an eagle, sometimes referred to as the "thinker's thinker." Whenever the eagle is said to "think," it is in reality engaging in first-order reflections of thinking about thinking. We attribute two identifying characteristics to the eagle: its altitude of elevation h and its very own consciousness c. From here we build a grid:

	High h	Low h
High c	bird-eye view	actively figuring things out
Low c	leisurely thinking	complete ignorance

To paraphrase Maria's view under our newly-adopted language of the thinker and the thinker's thinker, we have the following:

Claim 2.2. ***Maria****: The thinker possesses eternal telepathy with the thinker's thinker. Alternatively, the thinker's thinker exists internally within the thinker.*

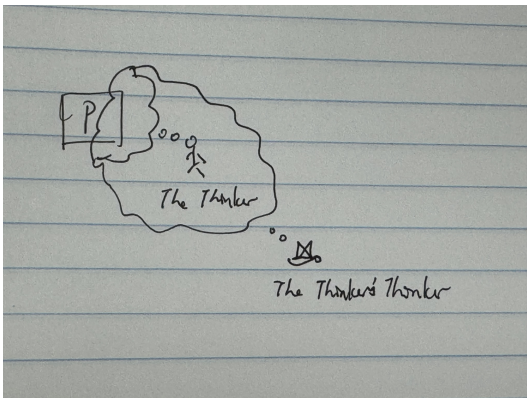
However, my experience suggests otherwise.

Claim 2.3. ***Lola****: The thinker is separate from the thinker's thinker. Each of them has free agency, and the thinker's thinker is metaphysically-ambiguous because we could only speak directly from the thinker's point of view. (Here we refer to the existence of a thinker's thinker as its conceptual image in the thinker's mind.) The thinker's perception of her thinker's thinker dual is an emergent property that will not arise until a sufficient amount of thinking.*

I shall explore my claim below. To reiterate, my perception of the thinker's thinker is an emergent property that will not arise until the thinker has either

(a) Thought sufficiently-long about p so as to be stuck; or

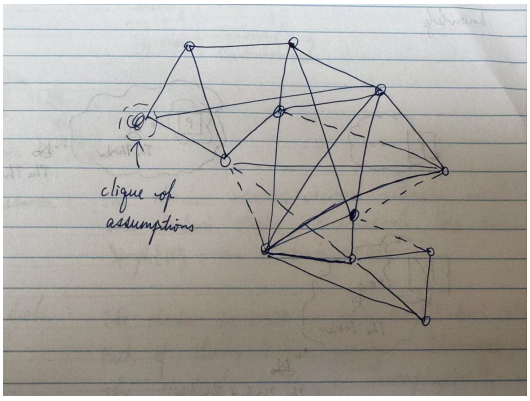
(b) achieved the epistemic belief that she has acquired all that needs to be known about p.



For my current model, thoughts of the thinker's thinker encapsulates (and hence arises from) the entirety of the thinker's thoughts p. Yet this consequently suggests that one wouldn't be able to think about thinking about p unless she has knowledge on p with sufficient measure. I shall explain what I mean by "knowledge with measure X."

2.2 Construction of Knowledge Structures

I tend to think of my knowledge structure as an infinite directed graph with vertices representing independent pieces of information.



At the graph's center lies a blurry (possibly-infinite) clique (completely-connected) of vertices representing all mathematical postulates and methodological assumptions subject to different areas of knowledge. For example, an assumption in science would be the PUN — principle of uniformity of nature, which underlies the scientific methodology of inductive reasoning. Another more easily-overlooked assumption may be that concerning human communication. We assume that the action of others are intentional, hence take efforts to make meaning -- propelling our study of the arts. We develop knowledge stemming from this clique, forming directed paths with one endpoint within the clique.

Definition 2.4. Two information vertices A and B are connected with bold edges if there exists at least one path from a vertex in the clique to A, and an individual is capable of deductively obtaining B by following that path.

For example, the integer axioms would be boldly connected with the information "1+1=2."

Definition 2.5. Two vertices A and B are connected with dashed edges if the knower cannot reason from A to B, but there exists at least one directed path in her knowledge framework (whether bold or dashed) from the cliques to A, and it is objectively true that there exists a bold path from the cliques to A and a bold edge from A to B, known by somebody "intellectually superior" than this individual of consideration.

This may be confusing, so let me highlight what I mean with a graph. Essentially, the dashed edges represent an element of trust in one's knowledge framework — a form of accepting knowledge claims from non-deductive reasoning. By "there exists a bold path" irrelevant to the individual's knowledge, I mean that there exists a way to deductively derive A solely from the set of assumptions coupled with the information from all the vertices our path has traversed, although the individual may not know how to. Later we will extend this definition to collective human knowledge, identifying a type of unperceivable bold line associated with "not knowing we don't know." Our definition of dashed lines thus allows us to quantify "by how much" we have relied on non-deductive trust in accepting a knowledge claim. We outline a graph algorithm that defines and calculates the measure of an information vertex P in an individual's knowledge framework. Let $T_G(P)$ denote the measure of P in someone's knowledge graph G.

(i) Identify all directed paths from the clique of assumptions to P.

(ii) For each such path, calculate
$$\rho = \frac{\#\text{bold edges}}{\text{path length}}.$$

(iii) Define $T_G(P) := \rho$.

This definition is not without problems. In particular, since we are taking ratios, a simple piece of knowledge (of distance $a \leq 50$ to the central clique) obtained from one or two elements of trust may have comparable measure to many mathematicians' understanding of Wile's proof to Fermat's Last Theorem – because they may have not taken the time to work through details of the proof; instead, they may have been satisfied by grasping a great picture and identifying new techniques they find valuable for their own research. Yet with this definition I may make clear an underlying assumption of my model regarding TaT.

Claim 2.6. ***Lola****: Thinking about thinking emerges only for a set of information $\{V_i\}$ for which*
$$\left(\sum_i T_G(V_i)\right)/i \geq \tau > 0, \tag{2.1}$$
where τ is a constant subject to individual thinkers.

My definition of a measure for information within a mental model reconciliates our disagreements regarding the intermittent occurrence of thinking about thinking. In particular, the factor τ could account for our difference in experience.
However, I must admit that my threshold for τ varies intensely across the subjects of thinking. In mathematics, the value probably approaches 1, while for philosophy of science τ may lie somewhere near $1/2$.
Next I would like to develop upon a structure for all knowledge - even those beyond human perception, which we apparently could only delve into conceptually. Assuming that each individual knowledge is a subgraph of the infinite graph of all knowledge K (where each individual information vertex in K corresponds to the same information as perceived by individuals), we could construct K' - a graph of perceivable knowledge - by taking the ideal union of all human knowledge graphs, replacing dashed edges with bold edges as long as there exists some knower whose knowledge graph consists of bold edges in the corresponding positions. We make some curious claims regarding K'.

(i) K' consists entirely of bold paths given the defintion of dashed paths for individual knowledge structure.

(ii) K' may contain illegal bold paths in a sense that no individual knower possesses a knowledge graph with an all-bold path leading to some information vertex P.

In a general framework for all knowledge, we adopt a different definition for dashed edges since individual trust (dashed lines in a single knower's knowledge framework) is no longer of our concern in a macroscopic setting.

Definition 2.7. Two information vertices A and B are connected with dashed edges if

(i) either A and B remain disconnected in any individual knowledge graph, or

(ii) at least one of A, B is nonexistent in any individual knowledge graph; yet

(iii) it is deductively true (but currently beyond human perception) that there could exist a bold path from the clique to B passing through A.

Notice that we could further classify vertices in $K\backslash K'$ into set N and Y:

- Y consists of information *known to be unkonwn*; while
- N consists of information *not known to be unknown*.

We may leave structuralization for a while and come back to it later - for in the next subsection I wish to explore the two classes of vertices.

2.3 Exploring the Known to be Unknown

In this section, I will journal about my classification of the unknown according to a status of meta-knowing. Precisely, how do we *know* that an information vertex is unknown? Should we expect to be able to define the intentionality of this set of vertices?

Consider, for example, the Riemann Hypothesis, a conjecture examined by the best mathematicians of our time despite remaining yet unsolved. As remarked by Terence Tao, *we do not have the appropriate tools to solve it.* For better context, we shall state the conjecture below:

Conjecture 2.8. *All non-trivial zeros of the Riemann-Zeta function*

$$\zeta(s) = \sum \frac{1}{n^s}$$

possess a real part of $\mathfrak{Re}(s) = \frac{1}{2}$.

We may only *speculate* that the conjecture is true – RH may or may not be an information vertex lying within the blurry regions of the graph, and we may only be certain of its existence once a deductive proof is revealed. However, this suggests our knowledge on RH may only consist of two states: an amorphous stage where even its existence as an information vertex remains under doubt, and a certain stage where RH migrates from a potential existence in $K\backslash K'$ to a determined existence in K'. Before examining this claim, we proceed to consult a few other instances of engaging with knowledge believed to be *known to be unknown*.

Scientific predictions may offer another example of things in the middle of known knowledge and that known to be unknown. Specifically, do unverified predictions in theoretical physics count as information vertices

(hence an integral component of the knowledge web)? It would seem rash for these predictions to be excluded from the vertices as soon as we consult the profound, rapid advancements of modern science, but it sounds equally impetuous for them to be included. Looking back on our original criteria for information vertices, we note that the clique of assumptions does not include postulates of scientific paradigms currently-in-use, but instead captures accepted reasoning in scientific methodology. We seem to have encountered a dilemma here: should our criteria for knowledge compromise rigor for purposes of applicatory success and potential advancements of the collective human knowledge, or should we never include claims of science in compliance to a love of certainty? Suppose we loosen our definitions a bit, and allow deductions within individual knowledge graphs to include postulates in the fundamental sciences. Clearly, the theorists went down this lane within their own knowledge graphs, treating their derived results as certainly known, instead of being *known to be unknown*. The theorists' knowledge graphs then create a discrepancy with our definition of the collective human knowledge, as we have specified that an information vertex would exist in K' as long as some human being has known it. Hence, if we were to accept the theorists' model of knowing, our vision of a collective human knowledge would need to be amended accordingly. Before we adopt this change, why should we accept the theorists' way of knowing at first hand? One may argue from induction that majorly acclaimed theoretical advancements are backed up by empirical success in later years, and since each individual knower's clique of assumptions includes the ampliative method of induction, she may accept the theorist's demarcation of unverified theoretical claims as knowledge. However, the very same clique of assumptions may have led to an argument in the opposite direction, if she consults a form of "pessimistic induction" by appealing to numerous falsified theories in the natural sciences. Notice, however, our knowledge framework treats a theoretical claim as a known information vertex as soon as it is validated by experimentation. This belief stems from the clique of assumptions' reliance on induction as a valid method. Yet our inclusion of these verified theoretical claims as information vertices introduce an element of uncertainty, implying possible vertex-deletions due to experimental falsification.
Our analysis of mathematical conjectures and scientific predictions have given rise to numerous (off-topic) questions, and they collectively imply the following claim on vertices known to be unknown:

Claim 2.9. *There exists no known to be unknown vertices conceivable by an individual knower.*

To justify Claim 2.9, we have consulted mathematics and the natural sciences – two areas of knowledge that seem to offer the greatest-possible certainty – yet an examination in either areas has given rise to an epistemic leap from an amorphous existence to some form of certainty.

In that case, then, do all known to be unknown claims reduce to unknown to be unknown ones? Both mathematicians and scientists would disagree, as such radical reduction strips away the epistemological value of the conjectures and theoretical claims themselves. In addition, there would be no immediate value attached to their work, especially to that of the scientist, as while the mathematician aims strategically at developing essential tools for a deductive proof, the scientist's paradigm-determined theoretical advancements rely on the empirical verification of experimentalists. We may introduce a hue of beliefs to rectify our current model's leaping nature, yet a justification of "how true a belief is" requires ample work.

2.4 Exploring the Unknown to be Unknown

A great part of this conversation was developed prior to my ascertaining the graph definition of a mental model, so some terminology may misalign. Nonetheless, I have tried my best to make changes that maintain coherence.

Lola: Is there a way for us to quantify the amount of knowledge that we don't know we don't know? By "we don't know we don't know" I refer to the knowledge that is not only epistemically barren but also conceptually unknown. For example, discoveries in theoretical physics like dark matter would uncover knowledge that was once not known to be unknown, as long as we agree the postulates that these deductions are based

upon are true. Certainly we cannot know what it is that we don't know we don't know, but can we estimate how many such things exist?

One way to quantify information may be to think of storing them in digital bytes, and since the amount of information stored for a single thing unknown to be unknown may vary violently, I shall clarify that this quantity should match the least amount of storage needed for one to comprehend the material while starting from their current knowledge base. For example, "$a^2 + b^2 = c^2$" must not qualify for an adequate amount of information for the Pythagorean theorem, as the equation per se does not explain the symbols. At the same time, a proof may not be necessary information for those familiar with rigorous proof structures and hence possess the ability to work it out.

"I think we should also highlight the scope of this question; does it ask for the unknown unknown knowledge for the collective humankind or just that for individuals? For now let's restrict our analysis to the limited, imperfect minds of individuals, because in talking about human knowledge as a whole we would also need to wrestle with concepts such as "collective knowledge" (i.e., is it possible for us to say that humankind knows X if none of the people p_i knows entirely X, but the union of all the knowledge x_i subject to each p_i covers X?).

Maria: (*unintentional response*) Was reading *There is no Antimemetics Division*, and I found it a fun read.

> Antimemetics are self-keeping secrets, yet they are nothing like the conventional passwords or taboos in reality, whose existence we have clear conceptions of. Nor are they transient, extensive concepts like dreams or complex mathematical equations, which cannot be conveyed to their fullest without referring to the entities themselves. Antimemetics are clearly-existent, as reported by empirical documentations of individuals who have just undergone an antimimetic experience. One could touch them, see them, and even stay with them – yet she will forget about them immediately afterwards. mnestics could counteract antimemetics by preventing or decelerating memory loss. There are four kinds of mnestics, labeled W-Z: "Class W mnestics allow somebody to perceive and continue having knowledge of antimemetics with memory enhancement on the side. Class X mnestics gives back awareness of antimemes or suppressed memories. Class Y gives somebody the ability to recall any memories gained during its effect. A single dose of a Class Z mnestic makes someone completely incapable of forgetting for their entire life. Class Z mnestics are fatal, causing death by seizure within a few hours." One's frequent, unspontaneous engagements with antimemetics could induce memory alterations of the rest of the population, causing an individual to be forgotten by people who happened to know her. qntm's SCP contribution *There is no Antimemetics Division* centers around the implied existence of an antimemetics division on Site 19, a division targeted to researching and investigating SCP-055 – a Keter-class object capable of memory erasure and hence possessing antimemetics properties. The story revolves around Marion, head of the antimemetics division, who has been forgotten by her cowokers precisely due to her engagements with the antimemetic SCP-055.

Lola: In some ways antimemetics resemble the realm of things we don't know we don't know, but there's something fundamentally different here: we don't know we don't know antimemetics despite having transiently known antimemetics.

I could see it, feel it, sense it ...but I may never remember it. I could enter a state of complete knowledge, yet exit with an exact form of former oblivion. I could know something for a while, and immediately unknow it; not even knowing I don't know it, nor knowing I have known it. I'm being brutally honest; SCP-055, oh what is that?

Fairytales aside, is there any real-life knowledge possessing antimemetic qualities at least in verisimilitude? I was considering the list of things the SCP foundation has compared antimemes to, but none of them function as a satisfactory real-world analogy.

> Think of any piece of information which you wouldn't share with anybody, like passwords, taboos and dirty secrets. Or any piece of information which would be difficult to share even if you tried: complex equations, very boring passages of text, large blocks of random numbers, and dreams...

Passwords: these are things we know we don't know. Ideal passwords are a form of "moknowledge" (a portmanteau of mono and knowledge), for which the password keeper knows the fullest of all available information, while the rest of the population *know* they may never know the contents of the password. Nonetheless, everybody knows not only the existence of a password, but also a guarantee of the limits beyond which passwords may not be exercised. In other words, we know that there are such things as passwords, and we know what passwords can and cannot do – so even if the contents of the password remains a forever black box, our conceptions of a password seem not very different from that of the cipher holder.

Things difficult to share: as the sharer ourselves, we know the information to its fullest, so in a sense the information bearer could be an antimeme. I do think it'd be easier to find an example by following the latter vein. Are there things in reality that intentionally prevent themselves from being perceived, shared, and spread by any human entity, despite their potential to participate in humans' sense-making process even amidst transient moments? While complex equations are difficult to be shared, they could nonetheless be partially perceived – and the transmitters themselves could potentially understand them to their fullest extent. Random gibberish or boring passages of text, unlike antimemes, evade comprehension not because of their intrinsic nature, but instead due to their universal lack of meaning.

Now consider dreams, something we could so vividly perceive in sleep that remains elusive amidst waking moments. Might we hope to extend this conception of dreams, and argue that in days when we wake up, not remembering to have dreamed, we may actually be engaging with some antimemetic experience? Yet to most people, the loosest degree of "dream evidence" would be a perception (upon waking) that they have dreamed despite their inability to recollect the dreams – so we may say forgotten dreams are quasi-antimemetics, but we retain some baseline levels of mnestics enabling us to recall the gist to have dreamed.

The problem of reality-based antimemetics lies precisely in our inability to know anything about its ontology. Unlike passwords or other forms of moknowledge, we may never hope to draw constructive analogies, and once we point to an entity in reality, we will have already perceived it and remembered it. Consequently, we may only hope to find an optimal approximation with maximal resistance to reality-induced mnestics. Dreams are the best I could think of.

2.5 On Thinking in Languages

Here we discuss the necessity to think in natural languages and interesting thinking phenomenons associated with multilingual dwellers.

Lola: Since you speak multiple languages, do you have a tendency to think in a specific language? Have you tried thinking about the same thing in different languages? How has that changed the result or your conclusions?

Maria: For some time now, I have primarily been thinking in English. A few years ago, my thought processes were more evenly split between English and Russian, depending on the language I was actively using that day. For example, if I were at home speaking Russian with my parents or reading in Russian, I was more likely to think in Russian. Nowadays, however, my thoughts are almost always in English. I recall being amused and delighted that in my dreams, I would seamlessly code-switch. For instance, if I were speaking to my grandmother in a dream, I would naturally use Russian.

I have experimented with actively thinking about the same topic in different languages to test the hypothesis that it might yield different insights. The most consistent effect I observed is primarily emotive. When thinking or speaking in Russian, I tend to express myself more emotively and am more inclined to "think politically," or at least that is how it feels to me. However, to an outside observer, the opposite might seem true.

Due to cultural differences in language use, Russian feels more emotive to me, but my English-influenced idiolect in Russian likely sounds more bookish or intellectualized to native Russian speakers, rather than emotive. In contrast, my use of English feels less emotive in comparison.

Lola: I could resonate with both the code-switching and the cross-language emotive influences you've captured, though the languages' emotive influence may have opposite for me. There are specific things that I've failed to think about while using a different language. A most common example being simple arithmetic: I could calculate the product of two three-digit whole numbers in my head with an internal-monologue in Chinese (coupled with symbolic numerals), but I always fail midway while trying to redo it in English. However, contrary to your experience, I think English is the only language where I could explore/express my emotions. It has also been my primary choice of internal monologue whilst I'm alone or while I am thinking to myself. I remember being driven to tears while reading a Chinese translation of Anton Chekhov's "Misery," which was on the suggested reading list from my school. The first few reflections (presented in the form of a natural language internal-monologue) that arose after I became conscious of my crying spawned in short English sentences that capture, without a conceivable trace of translation, the inherent kindness and cruel indifference all portrayed by scenes in the tale. What happened after these first conscious moments of crying was quite a schizophrenic experience – I tried asking myself "why are you crying?" in English once after once despite being unable to stop the tears, but as soon as I switch to Chinese the emotoins quickly wane out, possibly followed by some self-derogatory ridicules in similar forms as "how was that worth your tears?" I think this example also contrasts with the emotive influence you've described. But similarly, native speakers often find me emotionally-detached, despite my thinking that my emotions are most-strongly expressed in English. Or maybe this phenomenon hints at something at the intersection of both languages – my emotions may be the strongest when I am situated in some condition-specific billingual environment. I'd only be able to test this theory once I master a new language, and I do hope I'd be able to do so.

On the *arising of insights* claim, I do think switching between English and Mandarin helps me to better understand mathematics, though in either case my thinking process would consist of something more than natural languages. I've been fascinated by mathematical fallacies lately, and in reading through these fallacies I try to spot errors before they are commented upon by the authors. [Really interesting pastime, feels a lot like finding bugs in a program.] While sorting through the arguments I'd often need to rethink them quite a few times, and in these situations a bilingual way of thinking actually yields insights much quicker. However, the "much quicker" comparison is not grounded upon any deductive justification simply because we cannot control variables appropriately while reversing the order of code-switching: whilst thinking about X we will need to select an initial order of code-switching, and in rethinking X after reversing the order of code-switching our thought processes would already be influenced by the initial round of thinking. Nonetheless, I should try conducting another thinking-on-the-spot experiment. Here's a fallacy I've never thought through before:

Claim 2.10. *In any group of n people, all people are of the same age.*

Proof. We denote the statement of Claim 2.10 as $S(n)$ and argue by induction. First, $S(1)$ is true since a person is of the same age as herself. Next, we show that $S(k+1)$ is true by supposing the truth of $S(k)$. We aim to show that any pair of persons p, q share the same age in a set S of $k+1$ people. Notice that $S_1 = S\backslash\{p\}$ and $S_2 = S\backslash\{q\}$ are each a subset of size k, so all elements of S_i for $i \in \{1,2\}$ share the same age. Since $S_1 \cap S_2 \neq \varnothing$ and $q \in S_1$, $p \in S_2$, there must exist some $r \in S_1 \cap S_2$ of the same age as both p and q, thus concluding our fallacious proof. □

Below is a documentation of my internal monologue while scrutinizing the fallacy. While the proof has been reconstructed out of memory in English, I've read through it quite a few times in a blend of English and Mandarin before summoning enough courage to work "against it." All thoughts that are originally in Mandarin are italicized. So here we go:

(i) Consider the case when $n = 2$. Then S_1 and S_2 are disjoint so the argument fails for $S(2)$.

(ii) Wait *this* was *so* simple *how* come *I* let *it* gaslight *me* in *the* first *few* rounds *of* reading?

Okay, that was not a challenging-enough example. Let me try finding another fallacy: https://www.math.toronto.edu/mathnet/falseProofs/numbersDescribable.html.

Claim 2.11. *Every natural number can be completely and unambiguously identified in fourteen words or less.*

Notice that Claim 2.11 cannot be true; there are finitely-many words from the English repository but countably-infinite natural numbers. How, then, does the "proof" follow?

Proof. We employ a "proof by contradiction" by supposing that there exists a nonempty subset $S \subseteq \mathbb{N}$ for which every $s \in S$ cannot be completely and unambiguously identified. By the Well-ordering Principle, there exists some minimal element in S, say, s_0. Then we could call s_0 "the smallest natural number that cannot be unambiguously described in fourteen words or less," which is indeed a unique identifier. So S is in fact empty, and we have reached a desired contradiction. □

[Now we begin thinking.] Check the length of "the smallest natural number that cannot be unambiguously described in fourteen words or less." Yes, it's 14 word long. The argument seems to fallaciously apply to a broader class of claims; we may switch fourteen with some other number as long as the description stays consistent... What significance does "less" have? We could in fact delete the "less" and claim that every natural number could be described in exactly 14 words... We wopped. *I think we could also construct something similar for the negative integers. This is uninteresting though.* How is the implication from "s_0 is 'the smallest natural number that cannot be unambiguously described in fourteen words or less."' to the statement "s_0 is the smallest natural number that cannot be unambiguously described in fourteen words or less" justified? I don't know if they are manipulating syntax and semantics, or am I confusing simple things. Also, *what happens for the base case* $s = 1$*?* How can we justify that $s = 1$ has a unique and unambiguous identification? This also rings a bell on Russell's paradox, for as soon as a number is coined as "the smallest natural number that cannot be unambiguously described in fourteen words or less," it will have been uniquely identified. Okay, so our way of defining s_0 is logically-inconsistent. Yet does that imply s_0 does not exist? *Need* s_0 *to exist to attack the WOP proof.* There is such an s_0 and it is, not "its description is," the smallest natural number that cannot be unambiguously described in fourteen words or less. So the fallacious argument seems to be a... *play on syntax*? [Thinking ends.]

The original website provides a confirmation for my falsification, and it seems like my thinking has grasped the whole of it. https://www.math.toronto.edu/mathnet/plain/falseProofs/fallacies_46.html. I shall also remark that observations of thinking greatly distort my conventional thinking modes. Additional factors like the primarily-employed language in this document may have also altered conventional thinking patterns too.

2.6 The Arts as a Way of Thinking

Lola: In our one-on-one you mentioned that music, arts, and dance qualify as languages to you. Do you think a necessary criterion for language contains its users' ability to think in it?

Maria: Music, arts, and dance can be considered languages in the sense that they serve as tools for communication and can convey information. They can, at least partially, be abstracted into symbolic meanings. However, in terms of the ability to engage in conscious, logical thought, verbal natural languages remain the most suited. To the extent that one can "think" in art, music, or dance, it closely resembles "feeling the

world" through these mediums. This process can yield valuable information and insights, but such insights are often implicit, rooted in sensations. To analyze these insights more rigorously, it may be necessary to verbalize them in a natural language.

There are also different types of insights derived from these mediums. For example, one might learn a general truth from music, such as the profound emotional impact it can have, through the personal experience of deeply enjoying a song. This differs from using music to focus on a specific emotion, such as listening to a song known to evoke sadness to explore the multifaceted nature of that emotion. Additionally, one might listen to a series of songs with a particular chord progression to learn that progression, which represents a distinct type of learning and may generate a different form of thought.

In many cases, verbalization may not be necessary unless the insights need to be communicated to others. For instance, one can teach oneself to recognize a chord progression without knowing its formal name. However, noticing patterns and consciously applying that knowledge – for example, composing songs using preferred chord progressions – often requires some form of internal verbalization. Alternatively, recognizing a pattern may itself become language-like, as one implicitly tags different elements with the same abstract concept, even without using a socially accepted linguistic framework.

In essence, the type of thinking associated with music, dance, and similar mediums is deeply embodied and, by default, more implicit. This allows for an efficient form of complexity. For individuals with similar musical tastes, listening to a song together may convey a specific emotion more accurately and efficiently than describing it solely with words. The types of ideas that are most easily expressed vary across different modes or languages of communication.

This perspective aligns with the concept of enactivism, which posits that cognition arises through an organism's dynamic interaction with its environment, mediated by sensorimotor processes (https://en.wikipedia.org/wiki/Enactivism).

Lola: Re the "arts as tools for communication/information conveyance," I think one major problem with this take on the arts is the ambiguity of the mapping from the arts to their receptions. The agents that interpret the arts seem to *complete them*, and the intended messages of the creator may never be fully, let-alone correctly, grasped. [In fact, the creators' messages may vary or even wildy diverge from their initial intent through the acts of creation.] On the reception side, the mapping may be neither injective nor surjective – the underdeterminism in interpretations makes the mapping noninjective, and the existence of people who refuse to think [or interpret] music makes it further nonsurjective.

Re the "verbal language as the best tool for conscious, logical thought" claim: yes, I agree mostly with your statement, but I do hold that there are concepts, ideas, and sensations inexpressible by symbolic systems or mere natural languages. Like, in relation to your mentioning that thoughts derived from the arts require natural languages to be made explicit or interpretable in a desired way, we could imagine a scenario in which natural language about the arts comes before one really encounters the arts, as contrasted to cases where they see the arts, consult written-out literature on the arts, and then journal or structure their own language in alignment with their takeaways.

In the former case, natural languages seldom expand upon one's preexisting faculties. As contrasted to the illuminating effects of seeing arts for the first time, verbal descriptions of the art come nowhere near. This fact is showcased by our inability to communicate either emotionally-imbued music with the deaf, or the ambient world to the blind. Actually, I've been thinking about this for some time now; solving the problem of

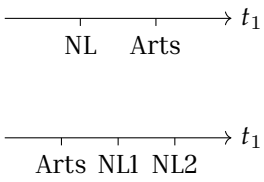
fully-verbalizing non-verbalized components of human perception may be an adequate first step to understanding the *thoughts* occurring in AGI processes. It's like switching between explanation and interpretation within Natural languages and the agents they interact with.

Re the "medium-specific insights" observation: I like your arguments. Let me paraphrase.

(i) We could perceive general truths from artistic mediums, i.e., insights about the *functionality* of certain disciplines. [e.g., music helps me appease my grief.]

(ii) We could then apply the function of these mediums to generate new, unlived perceptions.

(iii) Conversely, we may also inspect our existing perceptions by applying the functionality insights gained in (i).



3 A Step Back: Understanding Our Motivations

"Why did you do all this for me?' he asked. 'I don't deserve it. I've never done anything for you.' 'You have been my friend,' replied Charlotte. 'That in itself is a tremendous thing."

E.B. White, Charlotte's Web

Lola: Think I need a better rationale for why I'm asking you all these questions. Prior to SPARC, many of my friends avoided deep philosophical conversations, like they were steering clear of anything too introspective. Your insights stand out as especially valuable, so I'd love to keep these chats going indefinitely.

That said, I realize it's always been me leading with questions while you respond. Honestly, are you interested in what I ask? And in return, are there topics where you'd want my take? It's that mutual growth vibe: "*we are led to those who help us most to grow – if we let them, and we help them in return.*"

Maria: Appreciate it. Yes, I like your questions – they remind me of concept-building from scratch, like I do with one of my closest friends. On the asking asymmetry, I don't endorse you always being the one to ask, but I don't want to overpromise drastic changes just yet. A lot of my strand-holding comes from wanting to respond well. That's also why I initiate conversations less, at least over text, though I'm more like my old self in person. It's something I want to change, but I don't want to be too self-coercive about it.

Lola: I'm happy to keep being the interrogator in our chats – it pushes me to turn my half-baked thoughts into something structured, and our talks help me dig into the assumptions I didn't even realize I was carrying. The main reason I brought up the asymmetry is that I feel like I'm getting way more out of these conversations than you are, and I try to avoid situations in life where I'm just taking without giving back. I thought answering your questions might be a natural way to offer something in return, but if you're cool with just

answering mine, that's a perfect match.

By the way, I think holding strands is kind of an adorable habit, even if it looks like "ghosting" on the surface. For me, it's like letting a program run in the background while I'm out running errands or sorting through my days. Subconsciously mulling over those strands often leads to new insights about myself. That said, I lean on a life rule: if doing either A1 or A2 gets me to goal B, and I'm more emotionally attached to A1, I go with A1. So, I try to pick the best time to reply to strands—usually right after the "program quits unexpectedly" (like when I realize I'm not generating new insights) or when my enjoyment of thinking about them drops to about 75% of its peak. I'm still figuring out how to better spot that sweet spot.

On burnout, I totally get where you're coming from, though my threshold for overloading is probably lower than yours. Last year, I had a week where I was up until 4 or 5 a.m. every night, surviving on one or two hours of sleep, juggling four school research papers and two project writeups. That exhaustion wasn't just physical – it felt like someone yanked my soul out of me. I used to love the struggle, like unsticking myself from tough math problems, but that week, nothing sparked joy. After that, I started planning better. Even though I still don't always stick to my plans (like, oops, I should be writing my philosophy homework right now), sketching out deadlines gives me this illusion of control over my life, which somehow helps me enjoy things again. You seem pretty emotionally reserved, but if you ever hit a burnout and want to talk, I'm all ears.

Maria: Re the giving + taking, I care about this sort of symmetry a whole lot too... Think you might underappreciate how much I get from our interactions, even on the level of maths updates + seeing the kind of questions you're interested in and how you phrase it gives me the 'thinking about introspection from scratch' joy...

Relate both to having enjoyed doing difficult things and the spiritual depletion that burnout can cause... the week you described sounds hyper intense XD... Re setting deadlines as an illusion of control: reminds me of a thing I started to track at some point which is to what extent giving yourself constant strict deadlines which one gets used to not following through can create a chronic sense of failure/ self-disappointment even in worlds where the overall generator for control + discipline + difficulty still makes you perform well... Like I wonder how damaging internalising the cost of not achieving exactly the thing you promised yourself is, even if you made it consciously over-ambitious...

Had a bit of a wistful reaction to 'emotionally reserved'... I think I forget the extent to which I come across this way cause I tend to be fairly open in terms of sharing information about myself, especially when asked, but I do have a habit of e.g. sharing emotive things in ways that don't necessarily reveal how emotively I feel/felt about it. It's a family habit which I think of as a strategy for preserving openness amongst hyper sensitive people. My parents and I have always been very careful about 'burdening' each other with emotions precisely because we all know that we would have 'disproportional' reactions of care and concern towards each others 'problems' which would make it harder for us to concentrate on our work etc [for the same reason, we have a 'rule' in my family that if we ever have a bad argument we must resolve it before we go to sleep, so that no negative emotions seep into the next day], so we're honest with each other information wise, but try to be 'gentle' with the way we present the information... + separately, find it important to avoid 'reification' of negative things...

4 The Question-Answer Dichotomy

"I would rather have questions that can't be answered than answers that can't be questioned."

Richard Feynman

4.1 Opening Question

Once again we are discussing the art of conversations at a meta-level in Section 3. . . What processes are ongoing in our minds when we ask questions? How do they differ on the answerer side? I shall "ask questions about asking questions" to know more.

Lola: You got me thinking about a framework for quantifying the energy cost of conceptions. Intuitively, it does seem that asking questions is much easier than providing intelligible answers to these questions, but on what scale is this "easiness" measured? So far, my naive way of quantification involves a measure of *generative speed*, which essentially captures the respective rates at which a person verbalizes either her question or answers to her question. Major flaws of this quantification include

(i) it's subjectivity: say it person A verbalizes a question at a faster rate than person B; does that suggest the "required energy" of conceptions is, in turn, dependent on the individuals of reference? If we tailor energy consumption values to each individual's awareness, this quantification would be applicable to specific questions involving only one agent – a form of self-interrogation, as the lack of a universal scale makes cross-agent comparisons incommensurable. One may argue it suffices for us to compare energy levels in asking and answering for single persons to ensure metric consistency, but I could challenge this by alluding back to the giving-taking symmetry, of which I assume most intellectuals adhere to. Consider the situation where person α thinks of question Q with energy $E_\alpha(Q)$, and decides to ask person β, whom has previously thought of Q with energy $E_\beta(Q)$, where $E_\beta(Q) < E_\alpha(Q)$, supposing the energy metric is monotonically increasing. More questions arise from here. Is it necessary that β's original answer A_β to Q prior to α's asking must have also required less energy than α's conceived answer/working progress A_α to Q? Supposing that β is not in an energy-conserving state, would he consider himself not mirroring α's "giving attempts" if she merely repeats his original answers, although they are already much more developed than α's? (In case of a question with definitive answers as in math, β's answer may have addressed all it needs to be said, despite consuming lesser energy. Does β owe α anything on the giving-taking scale?) In the following table, blue texts are thoughts the respective agents have conceived yet failed to bring up.

Asking Space	Answering Space
$E(Q_\alpha) > E(Q_\beta)$	$E(A_\alpha) < E(A_\beta)$

(*reading this, I realize that I may be casted in the "low-efficiency" class of people, as what I wrote could be phrased so simply: if different people think of the same things with varying energy use, how do we compare broad categories of asking and answering?*)

(ii) and it's reliance on natural language, a strand I plan to challenge yet haven't yet responded to.

Another thing I wonder about is this: *how do we think or process information when a questioner provides a question we have never conceived of before?* I am trying to compare this with my thought processes while tackling questions that happen to emerge completely on my end, like many of my explorations in recreational mathematics. I think there is also an interesting difference in *conversion difficulty* of different types of questions. By "conversion difficulty" I wanted to refer to the extent of which an outer-imposed question could be internalized into one's own. At least for me, I'm unable to reason until I *understand the question* in a way where related questions that slightly divert from the original one emerge. For example, let me try drawing a question from a huge repository of inquiries in my Theory of Knowledge class (on the spot): "To what extent do you agree that doubt is central to the pursuit of knowledge?" Having just read this question, I proceed by recording the emerging internal dialogues:

(i) "doubt," "pursuit of knowledge."

(ii) [tries to graph doubt in the middle of a land called "pursuit of knowledge," and fails to grasp meaning of the question.]

(iii) When do I doubt? Do I know more after each instance of doubting?

(iv) The question seems unconcerned about the sufficiency of doubt in knowledge production. Is doubting necessary for us to know more?

(v) I think the question has much more relevance to me when I ask it this way: does the production of every piece of human knowledge involve an element of doubt?

(vi) Now I may proceed to brainstorm responses. . . I find myself alluding to specific areas of knowledge (e.g., math, the sciences, etc.) Might there be a conclusive framework for the role of doubt in the collective whole of all those fields?

This question would be considered of moderate difficulty in terms of conversion on my end. Next, I wish to compare this question with a math problem of similar length drawn just now from Mathematics Stack Exchange: (https://math.stackexchange.com/posts/5097204)
What Young diagram does the conjugate representation of $GL(n, \mathbb{C})$ correspond to?

(i) The question feels off.

(ii) Aren't young diagrams in correspondence with polynomial representations?

(iii) Conjugation is not a polynomial.

(iv) There are none.

I reasoned much quicker – notice how I was able to generate an intuition before consciously putting the properties of "young diagrams" in natural language. In the previous Theory of Knowledge question, however, I relied heavily on mental imageries and natural language. How does that come into play?

Yet another thing I plan to explore in the future is the *internal monologue* that goes on when I come back to previous questions I have thought about with, supposedly, fresh perspectives. I do not have the most perfect memory, so these processes oftentimes involve a blend of memory recall and active discovery.

5 Thinking about Thinking – Part II.

"Why do you think everything you think is everything everybody else thinks?"

George Saunders, *Pastoralia*

5.1 Opening Question

Our first round of thinking about thinking spawned a huge reservoir of thinking-specific questions. Beyond that, while reflecting on the process of meta-cognition, I seem to have spotted insights unique to "thinking about thinking about thinking." Please allow me to highlight a few of them below:

(i) The meta-cognitive chain of thinking about thinking could go on indefinitely, enabling us to "think about thinking," "think about thinking about thinking," "think about thinking about thinking about thinking," "think about thinking about thinking about thinkng about thinking about thinking," etc. For simplicity, let us use "think[i]" to refer to the i^{th} order of meta-cognition, in which the phrase "about thinking" is appended to the word "think" for i times. For example, "think" refers to a zeroth-order meta-cognition, while "think about thinking" refers to the first. Does "think[i]" yield fruitful results for every $i \geq 0$? Does there exist some i beyond which human cognition will not be able to perceive?

(ii) Whenever we say "think about thinking," we always append an entity to the thinking. In other words, we have always been "thinking about thinking about something." Could we think about thinking about *literal void*?

(iii) Philosophy of science is in itself a meta-cognitive process of thinking about thinking in science. By contrast to science, the cognitive towers of a philosophy of science seem much more accessible to a general public – the Theory of Knowledge course in the International Baccalaureate Curriculum includes a suggested chapter on the philosophy of science, and the scientific materials students appeal to are oftentimes beyond their technical grasp. How is this cognitive-trespassing possible? In general, how could we think about thinking about X, without having truly thought on X?

(iv) How do we make arguments? How do we make judgments? How could we justify non-deductive ways of persuasion, as commonly exemplified by arguments by analogy, contrast, or metaphors? If these ways of persuasion lie on no fair logical grounds, could we at least argue for their necessity?

(v) Poets often play on the word "think," creating scaffolding concatenations as "[somebody] thinks [someone else] thinks [yet another sombody] thinks X." When we communicate our internal meta-cognition, this epistemological chain is further complexified by an exchange of "about thinking about X" with the original "X." How shall we formalize this cross-individual meta-cognition?

5.2 Thinking about Thinking *void*

The mind is of or about things; intentionality is a primitive property of the mental. Before we proceed, I should justify its primitiveness by alluding to Dale Jacquette's famous thought experiment.

> Consider, he suggests, the phenomenology of intending in which an experimental subject refers to different aspects of a physical book (i.e., its colors, its contents, its formulation) all with the signifier "A," consciously switching the signified while maintaining the same signifier. One should notice two things form this engagement: one, she could use "A" in multiple referential ways without hindrance, and in each time of intending, an immediate decision has been made on her part instead of on reliance to inferential discoveries. Second, there is no essential accompaniment to her intending, thus making her thoughts directly intending the object. (i.e., nothing special before, nothing special after.) Since there are no psychological characterizations of intentionality, Jacquette concludes that intentionality is a primitive property of the mind.

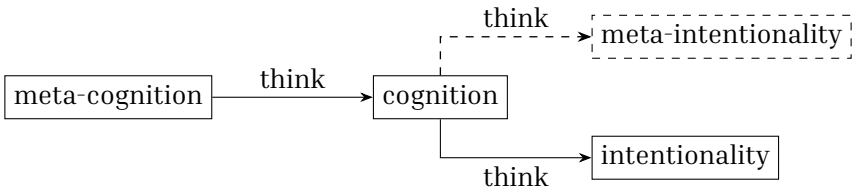
If we accept philosophers' claim about the necessity for the mind to possess intentions, we must then question the *intentionality of these intentions*. The mind is intentional in a sense that its thoughts are about objects or states of affairs. Certainly thinking is not an object, so could it be a state of affair? One might behold that it is also necessary for thinking about thinking to be about an object or state of affair, and we coin that to be a claim on *meta-intentionality*. Let's list out some intuitive arguments stemming from those ends.

(i) In previous instances of thinking about thinking, we have always been prompted to think[1] by specific objects, perceptions, emotions, or generally, what the philosophers call a state of affairs.

(ii) When we think, we must be thinking about something. Intentionality captures the representational necessity of the mental. Doesn't the meta-intentionality of think[1] follow from here? i.e., we need to always be thinking about *something*, so it naturally follows that whenever we think[1], thinking about thinking has to possess one further layer of intentionality since it is based on think[0].

Indeed, argument (ii) offers some fair grounds for us to doubt the possibility to think about thinking itself, but its logic does not deductively flow. In particular, not being able to "think" [without using the proposition about] does not imply our inability to "think about thinking" [void]. That is, we've wrapped the word "think" around "thinking." Indeed, this action does imply the post-wrapping phrase "thinking about thinking" relies on certain intentional quantity X; i.e., we must think about X and hence must only be able to think about thinking about X. However, this argument does not rule out the possibility for X to be the *action of thinking*, and as long as the act of thinking could be an intention in itself, the proposed argument would fail. Still, its failure offers not justification for the converse [that we could think about thinking *void*.]

Along these lines lies a radical challenge to meta-cognition: how is the boundary between zeroth and first order reflections defined? In practice, when is it just for us to say we are "thinking about thinking," and when are we merely "thinking?" Having accepted intentionality as a primitive property of the mental, we may attach any material object or state of affairs X to the end of "thinking," hence as soon as we accept "think[i]" to be a possible X for $i \in \mathbb{N}$, we will have gone in the way of supporting a reduction of meta-cognition to cognition. *If we could think about thinking void, then each instance of thinking about thinking is merely an occasion of thinking.* The acceptance of intentionality, coupled with a collapse of meta-intentionality would immediately lead to a failure of thinking about thinking.



Yet the implication from a denial of meta-intentionality to an inexistence of meta-cognition builds upon refutable assumptions. By arguing the contrapositive [no meta-intentionality → no meta-cognition], we presupposed that meta-cognition lies disjoint to cognition, yet it may very possibly be that meta-cognition is a subset of cognition. So we have hit a dead end and circled back to where we started. At this point, we do need a better definition for meta-cognition, or our colloquial term "thinking about thinking." Previously we have only offered an informal descriptive definition, so now we highlight a few essential properties of meta-cognition, hoping to elicit grounds for an extensive definition.

(i) Meta-cognitive processes arise only when non-meta-cognitive processes are activated. This does not imply meta-intentionality though, as occurrence may be disconnected with intention.

(ii) Meta-cognition is self-referential, it takes mental states as both inputs and outputs.

(iii) Meta-cognitive processes either form representations of mental states [which possess intentionality and hence imply meta-intentionality], or form a content-free representation of thinking [which consists plainly of the act of thinking]. This indetermination is at the hearth of our debate.

Upon further inspection, both (i) and (ii) seem doubtable. We present an original thought experiment:

> Suppose a "docile" AI agent with human-level intelligence believed capable of comprehending and thinking on itself is being prompted and starts thinking for the first time since its launch. Suppose (fantastically optimistically) that we humans have found the correct language to peek inside the AI black box. The

> prompt is engineered to be equivalent to what is meant by "think about thinking" in natural language. Once prompted, the docile agent starts "thinking about thinking," as its performance is calibrated to align with our will. Would that not undermine the necessity of non-meta-cognitive processes alongside meta-cognition as hinted at in (i)?

Yet in the AI agent thought experiment above, we may never ascertain whether non-meta thinking occurred alongside meta-cognitive reasoning. Moreover, the experiment is too vulnerable under the attack of Occam's Razor. It presupposes that the agent possesses no "knowledge" in a human sense, but is instead fluent in every form of meta-cognition; i.e. it knows how to think about thinking. How, then, has it been trained? Instead of developing this experiment any further, we present a more grounded alternative:

> Imagine a seasoned meditator Alice, who has practiced mindfulness for years. One evening, she sits down to meditate with the explicit intention of entering a state of complete mental silence. She is "not thinking," and has entered some void where conventional cognition has been suspended. Now, she starts noticing – a subtle awareness is present, watching over her non-thinking. Alice is not thinking about the sentence "I am not thinking"' there is no inner monologue or conceptual framing. Instead, it's a raw noticing of the mind's stillness, akin to the product of our senses before we label them with signifiers. This *noticing* is in itself meta-cognitive, and happens without any conventional thinking to scaffold it.

To push the experiment further: suppose we interrupt Alice mid-session, prompting her to report her experience. Alice describes it not through reconstructed thoughts, but as a direct intuition. Here the meta-awareness persists during the non-thinking state as the essence of the experience itself.