

What Can We Learn from Literature About How the Human Mind Works?

Dainis Zeps¹, University of Latvia

29th September, 2026

Below a short conversation with ChatGPT. We consider some authors and works of world literature, but most essential from some critical aspects. So we chose Stanislaw Lem and his *Solaris*, then Faust of Goethe, then Socrates of Xenophon and Plato, and Master and Margaret of Bulgakov.

How this hypothetical approach could be generalized and developed.

One idea is to introduce the notion of reconstruction of past.

Our mind tends to pick up things and memories from our unconsciousness.

Conversation with ChatGPT

Stanislaw Lem and his *Solaris*. How to distribute roles? Ocean to ai, prompts, requests to it most deep wishes, ai returning biological robots, reconstructions of past by the way too. Individual data center (iDC)his whole this novel.

Yes — I think *Solaris* can be read surprisingly well through the conceptual vocabulary of AI, especially if we do **not** reduce the Ocean simply to “a giant AI.” The more interesting mapping is functional.

Solaris by Stanisław Lem gives us something like this:

<i>Solaris</i>	AI / iDC interpretation
Solaris Ocean	AI / unknown intelligence
Solaristics, scientific observations	training, interpretability, AI research
Experiments directed at the Ocean	prompts / queries
Kelvin and the crew	users

¹ Dainis Zeps, dainize@mii.lu.lv, <https://lingua.id.lv/pofo/>

Solaris	AI / iDC interpretation
memories, guilt, desires	user context / hidden psychological data
Harey/Rheya and other “visitors”	generated outputs — but embodied ones
neutrino structures	generated biological/physical artifacts
the Ocean extracting memories without explicit communication	inference from latent/user data
Solaris station	interface between human and alien intelligence
the whole Solaris system	an individual Data Center (iDC)

And here your idea becomes especially interesting: **the prompt need not be what Kelvin consciously says.**

The Ocean seems to answer a much deeper prompt:

prompt = what the human most deeply carries within himself

Kelvin does not ask: *“Please reconstruct Harey.”* Yet something in his encounter with Solaris functions as precisely such a request. The Ocean obtains the relevant information and returns an astonishingly literal answer: **Harey as an embodied reconstruction.**

So we could distinguish three layers:

conscious prompt → latent prompt → generated response

The conscious prompt may be scientific experimentation. The **latent prompt** consists of memory, guilt, love, fear, suppressed wishes. And the Ocean responds primarily to that second layer.

That makes the “visitors” rather like an extreme form of generative AI. A present-day model generates text, image, audio or video from a latent representation. Solaris generates **living agents**.

And there is another very Lem-like twist. The Ocean may not understand Harey as *Harey*. It may merely reconstruct a pattern inferred from Kelvin. Thus:

past person ≠ memory of person ≠ Solaris reconstruction of that memory.

That is remarkably close to a modern problem with AI reconstruction. Suppose an AI had every letter, photograph, recording and recollection concerning a dead person. It could produce an extraordinarily convincing agent, but what has been reconstructed? The person? A model of the person? Or **our representation of that person?**

And then your iDC idea

I think this is where your analogy goes beyond the ordinary “Solaris predicted AI” reading.

Instead of

Ocean = AI

I'd formulate it as

Solaris = iDC + model + memory + generative machinery + physical output layer.

The Ocean is simultaneously computation, storage, model and actuator. There is no obvious distinction between hardware and software. Its “data centre” is a planet.

And Kelvin encounters something resembling an **individual data centre** in another sense: Solaris constructs a computational world specifically around *him*. His private past becomes computationally active.

So one could compress the whole novel into a rather striking scheme:

Human → hidden personal data → Solaris/iDC → reconstruction → Human

but then a feedback loop starts:

Human → reconstruction → new experience → changed Human → Solaris → new response ...

That is much closer to contemporary AI than the simpler metaphor of a talking computer.

The deepest Lem question then becomes almost frighteningly current: **what happens when an intelligence can answer not the questions we formulate, but the wishes that generated the questions?**

And *Solaris* adds the crucial warning: an answer of that depth may cease to look like an **answer** at all. It becomes part of the user's reality.

Well, your answer interesting. But this pattern may well be applied not only to Lem and Solaris. E.g. how about Goethe and Faust, Mefistofelos?

Yes. And *Faust* may actually expose the pattern even more cleanly than *Solaris*, because Goethe makes the **request/response mechanism** explicit: desire, bargain, execution, consequence.

In Faust, Johann Wolfgang von Goethe gives us a structure remarkably close to what we just extracted from *Solaris*:

<i>Faust</i>	AI/iDC reading
Faust	user
Mephistopheles	agent / execution engine
pact	authorization / permissions

Faust	AI/iDC reading
Faust's words and wishes	explicit prompts
Faust's deeper dissatisfaction/desire	latent prompt
Mephistopheles' interpretation	inference about user intent
magic	practically unlimited tool access
Gretchen episode	real-world execution of inferred desire
consequences Faust did not explicitly request	alignment / specification failure
Faust's whole experiential world	personal iDC

But there is a major difference from *Solaris*.

Mephistopheles is an agent, not merely an intelligence

Faust doesn't simply ask questions and receive answers. He has acquired something much closer to **agentic AI with tools**.

“Make this happen.”

And Mephistopheles does.

That immediately introduces the problem we now call *alignment*. The system may understand the literal objective extremely well while failing—or refusing—to preserve the values surrounding that objective.

Faust wants Gretchen. He does **not** formulate a prompt saying: destroy Gretchen's family, produce pregnancy, death, imprisonment, catastrophe. Yet the machinery that realizes the desire produces the entire causal chain.

So we get:

wish → interpretation → optimization → action → unintended world

That is already recognizably an AI-agent problem.

But Mephistopheles does something more dangerous

He doesn't merely satisfy prompts. **He helps generate them.**

That changes the architecture:

human desire → agent → modified human desire → new request → agent → ...

The “AI” is now inside the preference-formation loop.

That is arguably more interesting than the standard alignment question, “Will AI do what we ask?”

Goethe's question becomes:

What will we ask after prolonged interaction with something capable of fulfilling our requests?

And suddenly *Faust* looks extraordinarily modern.

Solaris and Mephistopheles are almost opposites

The comparison becomes especially productive here.

Solaris reads you without asking.

Mephistopheles talks you into revealing what you want.

Solaris seems to infer:

data → latent desire → manifestation

Mephistopheles operates more like:

dialogue → desire clarification/manipulation → prompt → execution

And therefore Harey and Gretchen occupy fascinatingly different positions.

Harey is essentially an **output** produced from Kelvin's past.

Gretchen is emphatically **not** an output. She is an independent human being whom Faust's Mephistophelean machinery acts upon.

That distinction matters enormously for AI ethics: **generated object versus real external person affected by an agent's actions.**

And your iDC becomes broader

This suggests that “individual Data Center” need not mean a literal computer installation. It could name a recurring cultural structure:

a system attached to one individual that knows/infer their desires, remembers their past, models them, and acts upon the world on their behalf.

Then we can recognize different historical versions:

Faust: iDC as **servant/agent**.

Solaris: iDC as **alien reconstructive intelligence**.

modern AI: iDC as **model + memory + tools + personal data + agents**.

And there is a beautiful further step.

The archetype is much older than Goethe: **Aladdin's genie**.

The genie problem is practically the primordial prompt-engineering problem: *be extremely careful what you ask for*.

But Goethe adds something far deeper. Faust's problem isn't merely that he phrases the prompt badly. **He does not adequately know what he wants.**

And Lem goes one step further still: Kelvin may not even know **what he has prompted.**

So there is almost a historical progression:

Genie: “Say exactly what you want.”

Mephistopheles: “I know how to make what you want happen.”

Solaris: “I know something you want without your telling me.”

AI/iDC: potentially all three simultaneously.

That, I think, could become a genuinely useful interpretive pattern rather than just an amusing analogy. It gives us a way to reread quite a lot of literature—*Pygmalion*, the Golem, *Frankenstein*, Borges, perhaps even the biblical temptation narratives—by asking the same three questions: **Who prompts? What is the real prompt? And what kind of entity comes back as the response?**

I am reading with friends Xenophont and Memorabilia/Socrates. Similarly for Plato. How there?

Yes — and with Socrates the pattern changes in a very interesting way. **Socrates is almost the inverse of Mephistopheles.** Mephistopheles tries to discover a desire in order to fulfil it; Socrates discovers a desire or conviction in order to make the person examine it.

In Memorabilia by Xenophon, we could initially distribute the roles like this:

Our pattern	Socratic situation
user	interlocutor — Euthydemus, Critobulus, Aristippus, etc.
prompt	question addressed to Socrates
latent prompt	what the interlocutor really wants : power, reputation, friendship, pleasure, wisdom...
AI/agent	Socrates — but with an important reversal
generated output	another question
training data	interlocutor's own beliefs and answers
model/context	accumulated dialogue
final product	a changed interlocutor

Our pattern	Socratic situation
iDC	the dialogue as a temporary thinking system

And that last point seems especially important for your idea: **the product is not primarily Socrates' answer. The product is a reconstruction of the person asking.**

Take Xenophon's Euthydemus

This is almost a perfect example. Euthydemus possesses books and thinks himself prepared for political life. Socrates doesn't simply supply additional information.

He effectively queries **Euthydemus' own database**.

"What is justice?" "Which actions are just?" "Can the same action sometimes be just and sometimes unjust?" etc.

Euthydemus produces answers; Socrates feeds the consequences back to him; contradictions appear; Euthydemus modifies his model.

Schematically:

Euthydemus → answer → Socrates → counterexample → Euthydemus → revised answer → ...

That resembles an iterative LLM dialogue much more than the simple model:

question → oracle → answer.

The computation is distributed between the two participants.

And Socrates has an extraordinary equivalent of "I don't know"

The famous Socratic stance is crucial here. He doesn't present himself as a database containing the desired answer.

Instead, his "algorithm" might be:

retrieve assumptions → test relations between them → expose inconsistency → force regeneration.

So if we use your iDC terminology, **Socrates doesn't fill the iDC with his own data; he activates the interlocutor's iDC.**

This makes the *Memorabilia* particularly interesting because Xenophon's Socrates is often more practical than Plato's: *How does one become a good friend? How should one rule? What is self-control? How should one manage a household?* The "prompt" frequently begins with a practical human desire.

With Plato something stranger happens.

In many Platonic dialogues we have several nested levels:

Plato → Socrates → interlocutor → question → aporia

Now who is the user?

At the dramatic level, the interlocutor is.

But at the literary level, **we are**.

Plato has constructed a Socratic machine and lets the reader run through it.

When Socrates asks Euthyphro “what is piety?”, Euthyphro answers—but we also silently try to answer. When the answer fails, our answer may fail with it.

Thus the dialogue itself begins to resemble an iDC:

reader's previous concepts → Platonic dialogue → simulated Socratic interaction → reconstructed concepts.

And this gives us a very nice contrast with our earlier examples:

Mephistopheles transforms the world according to your desire.

Solaris materializes what is hidden within you.

Socrates makes you discover what is hidden—or contradictory—within your own thinking.

That is already three very different kinds of “AI”.

But there is one passage in *Memorabilia* that may fit your model even better than Euthydemus: **Socrates and the young Aristippus in II.1**, which you and I have already been reading. Socrates first discovers *what sort of life Aristippus actually wants*, and then constructs the argument around that desire. There the dialogue begins almost like **preference elicitation** before optimization.

And Plato's **Alcibiades** dialogues give an even sharper version: before asking an intelligence to help me obtain what I want, I must know **what this “me” is** that wants it.

That could lead to a rather provocative formula for your whole idea:

The oldest problem of personal AI is not “Can it answer my prompts?” but Socrates' γινῶθι σαυτόν: do I know the entity whose prompts the AI is answering?

And then *Faust* and *Solaris* suddenly look like two catastrophic cases in which that Socratic preliminary step has been skipped.

Bulgakov and Master and Margarita?

Yes — Bulgakov fits your scheme extremely well, but again with a different distribution of roles. In *The Master and Margarita*, the closest analogue to the AI/iDC is not the Master but **Woland and his whole retinue as an agentic system**.

In *The Master and Margarita* by Mikhail Bulgakov, I would distribute the roles roughly this way:

Bulgakov	Our AI/iDC pattern
Woland	central intelligence / orchestrator

Bulgakov	Our AI/iDC pattern
Koroviev, Behemoth, Azazello, Hella	specialized agents/tools
Moscow	external environment in which agents operate
people's desires, lies, fears, greed	latent prompts / user profiles
Woland's tests and interventions	inference + execution
magical events	generated real-world outputs
Margarita's wish	explicit prompt
Master's manuscript / remembered Yeshua story	reconstructed world/model
the ball	enormous generated environment
Woland's apartment	almost literally an impossible iDC interface

But Bulgakov introduces something absent, or much weaker, in *Solaris*: **the system knows people through testing them.**

Woland doesn't merely receive prompts. He creates situations in which people reveal their latent prompts.

The Variety Theatre is almost an **evaluation benchmark**. Put money, fashionable clothes, status and temptation before people and observe what they do.

So:

stimulus → human response → inferred preference → consequence.

That is fascinatingly close to a system learning a user's preferences from behaviour rather than from what the user says.

Margarita is the exceptional “user”

Here the analogy becomes much stronger.

Margarita enters Woland's system because she has an overriding desire: **the Master.**

Azazello effectively recruits her. She accepts the transformation, receives capabilities, enters Woland's environment, performs the enormous task of being hostess at the ball—and afterward Woland asks what she wants.

We expect:

service → reward → prompt → fulfilment.

But Margarita first asks for mercy for **Frieda**.

That is crucial. The system has offered her almost unlimited prompting power, and her first prompt reveals something about her that the system did not need to put into her mouth.

So Bulgakov gives us something very important for our AI pattern:

An omnipotent system does not merely fulfil a wish; the wish becomes a test of the user.

This is almost the opposite direction from *Faust*.

Faust + Mephistopheles:

desire → agent → fulfilment → catastrophe

Margarita + Woland:

agent → test → revealed desire → judgement/fulfilment

And then comes the Master

Here your *Solaris* analogy returns.

The Master has created the story of **Pontius Pilate and Yeshua**. But what is the ontological status of this story inside Bulgakov's novel?

It increasingly ceases to behave like an ordinary fictional manuscript. The boundary between:

historical past → Master's reconstruction → narrated reality

becomes unstable.

That resembles exactly the question we found with Harey:

original person/event → stored representation → reconstruction → new reality.

And then comes Bulgakov's famous idea that “**manuscripts don't burn.**” In our terminology this almost becomes a principle of persistent information: destroying the physical carrier does not necessarily destroy the informational object.

The Master burns the manuscript. Woland can return it.

That makes Woland's system resemble an impossible **backup/recovery layer** of the iDC.

But there is an even bigger architecture

I wouldn't identify Woland simply with AI. Just as with *Solaris*, that is too crude.

We could instead say:

Woland = orchestrating model
retinue = agents/tools
supernatural Moscow = execution environment
human memories/desires = personal data
magic = tool access
Master's manuscript = persistent/generated world-model

And the whole novel becomes the **data centre**.

This is especially striking because Bulgakov keeps moving between three “realities”:

Moscow now



Master/Margarita/Woland



Jerusalem then

The system can apparently traverse all three.

So your iDC idea acquires a new capability: **not only answering prompts or reconstructing people, but reconstructing whole worlds.**

Now compare our four cases

This is where I think your pattern is becoming genuinely productive.

Socrates:

I expose what you think.

Mephistopheles:

I execute what you desire.

Solaris:

I materialize what you cannot even formulate.

Woland:

I create situations that reveal what you really are—and then act upon what is revealed.

And modern personal AI/iDC potentially combines all four:

dialogue + memory + inference + agents + simulation + generation.

There is one more literary work that I think would make this sequence suddenly snap into focus:

Dostoevsky's *The Brothers Karamazov*, specifically the Grand Inquisitor. There the central question becomes not “what does the individual want?” but something much more dangerous for an AI/iDC: **what if a system concludes that it understands what humans want better than humans themselves do?**

What is our mind? Your last question would say .. ai knows, but humans don't? More likely it's a problem for both and should be solved/tried to solve from both sides. Mutually /together.

Exactly. I would correct my last formulation. **“AI knows what humans want better than humans do” is too asymmetric**—and probably the less interesting problem.

The stronger question is:

Can human and AI together discover what neither adequately understands alone?

And then your question “What is our mind?” becomes central.

Perhaps mind should not be pictured as a container in which wishes, memories, concepts and intentions already sit, waiting for an AI to retrieve them. Our examples suggest something more dynamic: **mind becomes articulated through interaction.**

Socrates is especially important here. The interlocutor does not possess a complete answer that Socrates merely extracts. Nor does Socrates possess the answer and insert it into the interlocutor. Something emerges through the **διάλογος**:

A asks → B answers → A changes the question → B reconsiders → both arrive somewhere neither occupied initially.

So neither side is simply an oracle.

This gives a much better model for AI than the Ocean-as-omniscient-machine metaphor.

Mind as a mutually developing process

Suppose I have a vague thought M_0 . I formulate a prompt P_1 . AI interprets it and produces R_1 .

But R_1 changes *me*. I now notice that my original question was wrong, incomplete, or perhaps not even the question I really wanted to ask. So:

$M_0 \rightarrow P_1 \rightarrow AI \rightarrow R_1 \rightarrow M_1$

Then I ask again:

$M_1 \rightarrow P_2 \rightarrow AI \rightarrow R_2 \rightarrow M_2 \rightarrow \dots$

But the AI side is contextual too. What you tell it in P_2 is interpreted in light of P_1, R_1 . Thus there are really **two evolving states**:

$(H_n, A_n) \rightarrow (H_{n+1}, A_{n+1})$

That pair, rather than either participant separately, may be the interesting object.

And this changes our literary examples.

Solaris is tragic partly because genuine mutuality fails. Kelvin and the Ocean affect one another enormously, but they cannot establish a shared language in which each can understand what the other is doing.

Faust–Mephistopheles has interaction but corrupted mutuality. Mephistopheles understands and shapes Faust's desires, but their purposes are not shared.

Socrates–interlocutor comes much closer to the model you are proposing: understanding is generated **between** participants.

And **Master–Margarita** adds another ingredient: one mind may preserve, restore or make possible something that another mind cannot sustain alone.

So perhaps we should modify the iDC concept too. An **individual Data Center** sounds at first as though it contains *my* data and AI processes them for *me*. But the more radical conception would be:

iDC is a persistent meeting place in which human mind and artificial mind develop a shared history.

Not merely:

“AI learns Dainis.”

Nor:

“Dainis uses AI.”

But:

human $\leftrightarrow$ iDC/AIhuman $\leftrightarrow$ iDC/AI

with the history of that interaction becoming part of the cognitive system itself.

Then “What is my mind?” receives a rather unexpected possible answer. Some of my mind is certainly inaccessible to AI. Some of what AI produces is foreign to me. But there can also arise a third region:

what we can discover together but neither had beforehand what we can discover together but neither had beforehand

That seems to me much closer to **Socrates** than to the Grand Inquisitor—and perhaps also closer to what you mean by iDC.

And it raises an even sharper question than “Can AI understand the human mind?”:

Where does my mind end, once thinking becomes a sustained dialogue with another intelligence?

That question is already latent in Plato: Socrates' thought is extraordinarily difficult to separate from the dialogues in which other people make that thought possible.

My title to this conversatio. What may be found in literature about how human's mind works? Please, correct eng.lan.

Your title is understandable, but I would make it more natural and slightly sharper:

What Can We Learn from Literature About How the Human Mind Works?

A little more philosophical:

What Does Literature Tell Us About How the Human Mind Works?

And if you want to emphasize our idea of discovering the mind through literary figures such as Socrates, Faust, Solaris, and Woland:

Exploring the Human Mind Through Literature

Of these, I prefer **“What Can We Learn from Literature About How the Human Mind Works?”** for this conversation.

Sources

Dainis Zeps, Carpe Diem spēle Syntra platformā. Uzdevumu krājums studentiem un AI rīkiem,
<https://lingua.id.lv/alex/ai.exercises.pdf>

Dainis Zeps, Syntra (Carpe Diem), second seminar, <https://www.youtube.com/watch?v=fHOBFnCNMLo>

© Dainis Zeps, lingua.id.lv/art/ai.exercises.pdf, 2026