Goals of this doc

- To flesh out, between us (consortium members), a better shared understanding of the mechanics of the thing we're trying to build (including the interactions with all kinds of agents and the world)
- To clarify what parts we still don't have a good idea about
- To get a clearer scope of what the PoC needs to be and what else we need to prove our theory(ies) of change

Color-coding

- Red: We don't even have a clear theoretical the idea
- Orange: We have a clear idea, but don't have a PoC
- Yellow: We have built a PoC (or this exists already), but aren't satisfied with the validity
- White: We're satisfied with the idea's validity

As simple of a description as we can provide

The ultimate goal of this project is to create a shared world model network, the Gaia Network. It's like Wikipedia, but instead of *articles* and *links* between them, we have (causal, probabilistic) *models* of specific aspects or subsystems of the world, and *relationships* of containment, abstraction and communication between the models. Agents (humans, AIs, or entities) will use this model network to make inferences and predictions about the future, about parts of the world that they can't observe, and especially about "what-if" scenarios.

Whereas each Wikipedia article is "about" a specific topic (defined by the article's title and potentially disambiguated in the article text), each Gaia model is "about" a specific *target*, a system or a set of systems in the world (this "aboutness" is defined by a machine-readable *semantic context*, or just *context*[1]). The context specifies how the model is "wired" to the target: which input data streams or sensors are accessible/relevant to the model; which actuators are controlled by the model; and how both of these are connected to the model's internal variables (the "local ontology"). A node in the Gaia network is uniquely identified by its semantic context (using content-addressable identifiers) and must always include a model, even if that model is trivial. A context may be concrete (addressing a target we identify as a single system in the world, ex: "this square of land that today is used as a farm", "that person") or abstract

[1] This conforms with the formalisation of context in Fields, C., & Glazebrook, J. F. (2022). Information flow in context-dependent hierarchical Bayesian inference. Journal of Experimental & Theoretical Artificial Intelligence, 34(1), 111–142. https://doi.org/10.1080/0952813X.2020.1836034

(addressing a set of systems bound by some commonality, ex: “all farms”, “all people”), but this is just a fuzzy distinction to help conceptually (ex: when discussing hierarchies of models).

Like articles in Wikipedia, models in the Gaia Network are contributed by its users; these contributions are validated by the users themselves. Unlike Wikipedia, which is effectively a unitary knowledge base, knowledge in the Gaia Network is pluralistic: agents can host Gaia nodes containing their own models of the domains that they know best (for which they have data). Instead of forcing an agreement on a single “global” version of the truth like Wikipedia, the Gaia Network allows for a plurality of models for the same domain to coexist and compete with each other (agents choose which model makes the most accurate predictions for them), while being able to interoperate between contexts.

The Gaia Protocol is the set of operations that agents use to establish the Gaia Network. These operations include publishing nodes and their contents (intended contexts, models, plans and decisions), and querying the network for estimated outcomes of potential plans, model discovery, updating shared models, etc. It is roughly the equivalent of the “Wiki” content management and governance software that powers Wikipedia (and Wiktionary, and many other private Wikis).

Alt take: We already have a global, bottom-up super-repository of knowledge: the Internet. The Gaia Protocol is a layer on top of the Internet (and, perhaps, trust spanning layer for decentralized identity and dataset/datafeed governance/management?) that:

a) enables it to represent and dynamically compute beliefs and counterfactuals via a common language of Bayesian (causal) models;
b) grounds these beliefs in intersubjective claims and commitments provided by the users of the system, via a shared, Bayesian-coherent accounting system and knowledge economy.

The Gaia Network is to the Gaia Protocol as the Web is to HTTP/S. Everything else just follows from this.

The theory(ies) of change

Because of the lack of shared, comprehensive and reliable models, people make less-than-ideal decisions for themselves and for their collectives, and miss opportunities to coordinate on positive-sum strategies. This compounds exponentially with the number of decision-makers and the complexity and interconnectedness of the world. And this problem applies even more strongly for AI decision-makers, which are blind to their broader context (they don’t know what they don’t know, so they routinely make confidently wrong inferences and decisions).

Hence, our “master theory of change” is:

1. Design the Network architecture and incentives in such a way that it grows and, as it does so, it will also converge on a truthful world model and be resilient
2. Get it to grow, monitoring truthfulness and resilience and adjusting/governing as it goes
3. Advocate for increased reliance on the Network for both human and AI decision-making, based on empirical evidence of truthfulness and resilience

“User agents” vs Ents

Primarily, when we talk about users interacting with Gaia, we’re talking about humans acting on their own behalf or of organizations (corporations, governments). These could also be software agents (perhaps driven by LLMs).

Any user can spin up a Gaia agent, just like an HTTP server. However, they have to share useful information to other users/nodes in order to accrue trust, and they have to accrue trust in order to be widely used. Trust is related to but not the same as FERN, so it’s more of a meritocracy than a plutocracy: see below.

Anyone can just use Gaia, but many nodes will charge FERN for usage.

As the contents of the Gaia Network may/will include fitted ActInf models of individual things in the world (ex: people, corporations, farms, power plants…), these “dTwins” can be hooked into controls to make them full-fledged representatives for those things. These are what we call Ents. Fangorn is supposed to have native support for them. Ents can also be user agents in the above sense.

If the Network has rich enough model content for a domain, it becomes economical to give an external agent “live knowledge” by just querying the network (connecting with the hive mind), instead of training it on the domain data/knowledge. In the longer term, we envision that arbitrarily complex general-purpose agents will be (automatically) designed in this way, giving rise to “architectures of shared intelligence” on the Network.

Is this a crypto thing?

We want to pragmatically borrow/reuse the simple, good things from web3/crypto, namely global state management through blockchains (more generally, local (peer-to-peer) state management through Merkle-DAGs, as implemented in DefraDB), and potentially some of the ongoing innovations in tech and economics being developed for DAOs.

TODO: check what can https://github.com/subconsciousnetwork/noosphere offer

Value of information (the accounting system)

Sharing data and models is key to achieving economies of scale and scope in the network. Yet data and models are valued differently by different agents, and in the aggregate, some data and

models are simply more valuable to the network as a whole. This and the below section will discuss how these facts translate into a coherent accounting system and a knowledge economy that incentivizes convergence on high-quality knowledge well-distributed across the network.

For the following, assume fully public models and data. We'll talk about privacy later.

Every Gaia node *i* contains a Bayesian model *M* of some thing or aspect, fitted to some data *D* (which can be coming from other nodes or from users; these can include not just hard data but also "soft observations", meaning posterior beliefs.) Hence it can compute its VFE of current beliefs. Hence, whenever it receives new data *D'* and performs updating, it will calculate a VFE delta. This reduction in VFE (or increase in ELBO), measured in bits or nats, is this node's "subjective value" of D'.

If M is a proper active inference model, the node will also compute its EFE under various policies, and in particular *G**, the EFE under its optimal (intended) policy. So in the same way, new data will result in an EFE delta, also measured in nats: this node's "subjective expected future value" of D'. This decomposes into epistemic and pragmatic values, as usual. (This is exactly what the "applied information economics" school defines as the value of information.)

The sum of those two deltas is the subjective value of D' for i.

Then D' gets added to the overall dataset D, rinse and repeat.

The same applies when the node receives a model update M'; it performs Bayesian model comparison between M and M + M' given D, and calculates the same values as above. This is the subjective value of M' for i.

The value of information (either models or data) can be negative. In the case of data, these are "outliers"; considering them in the data set increases uncertainty and may make the model less able to do the right thing. A simple node can ignore outliers; a more sophisticated one will include a measurement model where information from outliers gets absorbed by source-specific precision parameters. Meaning, it's actually turning a negative into a positive by learning which sources not to trust. If the node's current model doesn't include measurement, the node can keep D' in its "back pocket" and re-add it once it does, keeping a form of "optionality".

Similarly, if the node has endogenous model selection, it can simply ignore a model update M' with negative value. It would probably still benefit from regularly revisiting M' as it receives more data which may justify adopting it. A still more sophisticated node can perform Bayesian model averaging and keep both M and M + M' as alternate models, with the latter having a low posterior weight until and unless new data comes along that justifies increasing it.

Note that in the above we have a strictly "online", time-sequenced inflow of information. This means that the value of information for a node is intrinsically tied to the specific moment in time

when it receives it. This is unlike “offline” approaches like leave-one-out where the node can imagine what it would have been like to not know something.

The above means every node keeps a private ledger of the value of every piece of information it has ever received, attributed to its source. This is the most basic/fundamental version of what we call the accounting system. It corresponds to the “deprival value” system in accounting. As we’ll see, a more familiar “mark-to-market” accounting system is also possible, but only once we build an intersubjective (market) value, resulting in a knowledge economy.

The knowledge economy

This is how the subjective value of information gets turned into an intersubjective quantity (FERN). The system is basically automated market mechanics (as commonplace in trading bots, ads marketplaces, etc), but using active inference as the valuation mechanism. Here’s how it could work: (The examples below are all with data, but the same applies for models.)

Each node i can always compute, for every source j and for every observation modality m it supports, its willingness to pay (WTP), the maximum price in FERN it would pay for data about m from j. This is the expectation (weighted by its current full joint posterior distribution) of the value of receiving an arbitrary additional data point about m from j.

A node can also compute a minimum offer price. For instance, it can amortize its own data costs, its compute costs, etc. It can also set personalized offers by computing the expected value of information offered to a peer, based on its beliefs about the current beliefs of the peer. Finally, it can decide to have a zero offer price, if it wants to be “altruistic” and share information as a public good.

When i receives an offer of data D from a source j, they negotiate the price P between i’s WTP and j’s minimum offer price. An atomic transaction happens where j shares D with i, while i automatically attributes P nats to j on a global FERN ledger (distinct from its private ledger). This is essentially how Ocean Protocol works.

In general, i’s post-facto subjective valuation of D will be different from the ex-ante price it agreed to pay. This is working as intended. If a seller j systematically provides bad (negatively valued) information, the buyers’ source reliability models will learn that j is not to be trusted; this factors in automatically into their WTP.

Additionally, as information that certain sources aren’t reliable is a public good, there can be a public good funding scheme where buyers are incentivized to share it, or they can altruistically share it for free. This forms a probabilistic analog of a web of trust. As the information about reliability percolates through the network, it reduces the incentives for low-quality sellers.

This can also support some kind of staking mechanism for added skin-in-the-game, especially in cases where evidence is unavailable/scarce, unreliable, or decisions have high stakes relative to the cost of publishing information. Ocean Protocol also has an implementation of this, but it's very limited and rife with crypto liquidity/pricing issues.

Both consumers and sources of information can be simple passthrough nodes; for instance, an IoT sensor can be attached to a trivial ActInf model that just treats the sensor data as a sharp (one-hot) observation and publishes/sells it on the network. And a human buyer can interface with a trivial ActInf model that just "chooses to" buy whatever the buyer tells it to, at the selected prices. But the value comes with more sophisticated models.

Model generality, federated inference, and science

The fact that agents value models based on their usefulness contingent on their own contexts means that intersubjective model value is intrinsically tied to generality. In particular, if an agent buys a model and then modifies it to fit its own context (ie, by reinterpreting inputs and outputs), it is under no obligation to share the value it derives from the modified version to the source, only the value from the original.

The extreme case: say node i has created a model of a specific system ("Alice's farm" in Wisconsin), and it's specified in such a way that it's not generalizable at all. (Let's say, it used ML to create a highly predictive/useful model for managing that farm, but which knows nothing about other farms.) Then it can try to share the model on the network, but no other agent will want to pay for it. Worse: if it shares the model for free, other agents might take the model, find ways to rewire it to their own contexts (perhaps trivially), and get a ton of value out of it, which will not be shared with i at all.

This means agents are incentivized to specify their models in as general terms as possible – far beyond the bounds of what their local contexts would ever care about – in order to have any value in the model market. This inevitably leads to hierarchical models, where higher layers model beliefs about very general facts (ex: average growth rates of all corn anywhere), matched to a global ontology; while lower layers model beliefs about very specific ones, potentially referring only to a local ontology (ex: growth rate of this particular corn seed lineage in this particular hectare assuming a wet, hot year, this particular cover crop, etc.).

Obviously, if an agent only has access to information about its own context, it won't do very well at generalization: the two layers will carry the same information. So what it must do is acquire information from other agents with similar enough contexts to justify knowledge transfer. Say j is a node for another farm ("Bob's farm" in New Zealand). Assume their model *structures* are at least partially compatible at the higher layers: ex, both have variables representing beliefs about the average growth rates of all corn anywhere, associated to the same names in the ontology. We typically call these variables "global" to distinguish them from "local" (target-specific). (The nodes may have independently arrived at it (convergent evolution), gotten it from the same source (more on that later), or j may have bought it from i as discussed above.) Then, let's say i

receives new local data; as part of the usual Bayesian updating, all the global beliefs are updated. This means i now has an updated marginal posterior for the global that incorporates information that no other node has, and it can share that posterior with j (as *data*, not a model!). j, in turn, wants to acquire i's global posterior because it can use it as a prior to better interpret its own local data. And of course, the same is true in the other direction: when j receives new data, it can share its global posterior with i, who wants to acquire it for the same reason. This means that i and j jointly learn the "true" posterior for the global variable and use it for their own ends (which can be radically different).

What we've just described is an implementation of partial pooling via federated inference *["FedBMR" in Fangorn]*. But another way to look at it is that the nodes are simultaneously performing science (systematically generalizing from local observations to widely applicable knowledge) and science-based decision-making (systematically instantiating that knowledge into priors for local decisions).

The strictly peer-to-peer inference scheme is guaranteed to converge to an approximation of the true global posteriors, but it may take a long time to do so (especially if nodes have reasons to not fully trust updates from their peers, more on which later). Additionally, it only works if a critical mass of nodes aligns on partially compatible global models, which, as we discussed, may happen in a purely market-based scheme, but perhaps only very slowly. Hence, it would likely be valuable to set up "science hubs" as independent Gaia nodes, both to serve reference global models for free, and to aggregate global posteriors from compatible nodes, sharing them back with the network as scientifically backed, up-to-date, cross-validated reference beliefs *["Entmoot" in Fangorn]*.

Epistemic vs pragmatic agents

Sensemaking/model alignment when observability is low

The delegation economy

EFE estimation is always done relative to a policy. Policies are distributions of future pathways over an action space. This action space is, strictly speaking, limited to an agent's direct *capabilities*, i.e., what variables it has control over. However, in a network of connected agents, an agent has control over further variables that represent its interactions with peers: communications, transactions, and commands. This opens the door for agents to model the capabilities and motivations of other agents, and to attempt to control/influence/motivate them in order to indirectly expand its own action space. We generally refer to this as *delegation*, just to emphasize that the key fact is indirectly achieving that which you can't indirectly do, but the term currently seems to have "top-down" connotations, whereas we also want the reader to think about lowly actions such as voting and paying for services using the same term. (*Transaction* is

typically used in economics, but it carries its own connotations of something… well, transactional [atomic] and symmetric.)

Humans can delegate quite easily: as a social species, we have strong priors about our peers' capabilities and *affordances* (how they can be motivated and influenced). How to do this in a network of heterogeneous agents which may have wildly different capabilities and affordances? Likely there will be a combination of:

- Agents must publish their self-models.
- There must be some degree of observability of the input-response pairs.

Both of those are straightforwardly supported by the Protocol and the knowledge economy, but will benefit from the creation of additional "skills marketplace" type services on top of it. But once those are in place, the usual machinery of decision-making by EFE minimization applies to decisions about what to delegate to whom, given one's "budget" of incentives and the space of potential agents with different capabilities and compatible affordances.

With relatively small additions, this setup can also support:

- *Commitment mechanisms,* where agents can reduce problems from delegation, such as moral hazard and simple uncertainty about agents' *actual* capabilities and affordances. (Note that even though Gaia agents always have introspection [knowledge of their own models], they might still have substantial uncertainty about what they can actually do and what they would actually choose when confronted with a given situation/incentive. This is particularly relevant in contexts where agents are sub-delegating to other agents, which of course is widespread in human social systems.)
- *Coordination* mechanisms, where policies involving multiple agents (and typically affecting multiple target systems) can be designed, committed to, and monitored.

Where do preference priors come from?

In order to calculate EFE, an agent also needs a preference prior over its state space. (If it just has flat priors, it will select actions purely on epistemic grounds.) The preference prior is often spoken about as equivalent to a utility function in decision theory, but it's best understood as a biased distribution of beliefs about the future – and this means that it can be updated, just as any other distribution. Indeed, the general idea in ActInf literature is that "the higher-order dynamics" of the system (up to and including the model design process) should inform the preference prior. This is a fundamentally different (and better) idea than RL concepts like self-play, which are fundamentally non-contextual. Unfortunately, we have very little practical prior art to lean on here: published ActInf models have fixed preference priors.

The Gaia Network implements preference updating as part of model sharing. For this to work, agents only need to have ActInf algorithms capable of counterfactual goal-setting: "what would the future look like if I acted according to these preferences?" The key idea is simply that certain

sets of preferences are instrumentally more rational, ie, they lead to a lower optimal EFE if pursued.

Satisficing/bounded rationality, bias helps preferential ideation and exploration/identify solutions
This does imply something like “algorithmic wisdom”, where

Where do models come from?

The convergence arguments in a nutshell

Since the Gaia protocol is very simple, a proof of existence is easy. However, the real question is: once it exists, will Gaia do what it says on the label? Will it be what we want it to be? Specifically, we want to argue that:

- Gaia will converge towards a truthful body of knowledge (the world model) that supports effective and appropriate decision safeguards;
- This world model will be resilient against adversarial behavior.

As we discussed above, formal guarantees are great when possible, but when they involve oversimplifications, they tend to lead to false certainty (the map/territory fallacy). So we ask the reader to first consider the following heuristic arguments.

Epistemic status: The extensive indirect evidence for this argument puts us at ease, but we’re working to obtain direct evidence: see the “Conclusion” section. As does the existence of two separate arguments.[2]

1. **The bottom-up argument:**

This argument follows two mutually reinforcing parts:

(a) To the extent that contamination can be filtered out or ignored, Gaia will converge towards a truthful body of knowledge;

(b) to the extent that Gaia has truthful knowledge, it will be economical to filter out or ignore arbitrarily large amounts of contamination. It is structurally the same argument that explains why open-source software works, and it's at heart an evolutionary process theory.

[2] These two arguments roughly correspond to the “low road” and “high road” to active inference in the book by Parr, Pezzulo and Friston.

Part (a) follows from the classical (Jaynes) view of Bayesian probability as the logic of science. If agents can trust any given state *s* of Gaia's knowledge as an approximate representation of the world, they can then use it as a prior for Bayesian updating and share their posteriors on the network, forming a new state *s'* that is likewise a progressively better representation of the world (and, in principle, this improvement could even be calculated in a consensual way as free energy reduction). In this way, all materially relevant[3] errors in the collective knowledge (whether originally coming from noisy observations or faulty models) will eventually be corrected by further observations, experiments and actions. The general case is that an agent will believe that the content of Gaia is a mixture of knowledge and contaminants, and will attempt to discriminate between the two using *trust models* (which may, for instance, describe a certain dataset's probability of being bogus as a function of knowledge about its source's identity or past behavior.) These, in turn, are just more models on the Gaia Network. So if Gaia is even moderately useful and trustworthy, it will be adopted and contributed to, which makes it more useful and trustworthy.

Part (b) hinges on both the append-only nature of the network's contents, and the intrinsic reversibility of Bayesian updating.[4] Even in the presence of concerted effort to contaminate the network with bad models (including conspiracies to give high weights to those models), these merely add noise to the network; they don't replace the genuine knowledge already accumulated. Through simple "revisitation" procedures, agents can roll back through content, sift the wheat from the chaff, giving higher weights to the high-quality models, and collaboratively downrating low-quality ones. This imposes an increasing hurdle for sources of low-quality models to contribute, such that, over time, the signal-to-noise ratio improves. Hence, the more contributors engage with the network – indeed, the greater the number and variety of erroneous models it ingests – the more robust it becomes against contaminants. This is analogous to the immune system in biology, where exposure to a variety of threats strengthens the system's ability to identify and neutralize harmful agents.

Furthermore, the economic aspect of part (b) becomes evident when considering the cost-benefit analysis of participating in and maintaining the Gaia Network. As the network

---

[3] The world materially here means: If there is knowledge that cannot be gleaned or corrected in this way – as some believe to be the case about the foundations of physics, cosmology, and other fields – it is not relevant to Gaia's purposes. Similarly, the fact that we may never know the truth about historical facts lost in time is irrelevant here.

[4] Reversibility is lost when an agent assigns a probability of zero or one to a belief, which effectively represents infinite confidence in its falsehood or truth, a "belief singularity" from which no evidence can remove it. Procedures for explicit Bayesian updating (or its approximations) are intrinsically safe from this mistake; however, this problem arises implicitly and routinely when, in the course of the modeling process, a model selects an excessively simplistic or rigid structure for the context at hand. By disallowing alternative hypotheses as an initial prior, an agent using such a model is effectively locked into such mistaken and unmovable assumptions straight out of the gate. This is one major reason why the Gaia Network's design emphasizes higher-order, context-aware metamodeling.

becomes more accurate and reliable, the relative cost of filtering out contamination decreases, while the value of accessing the world model increases. This creates a positive feedback loop, encouraging more participation and investment in the network, which in turn enhances its quality and trustworthiness.

2. The top-down argument

This argument claims there is an attractor pulling the world system towards ever greater amounts of shared sensemaking and coordination. Specifically:

1. Given the brute facts of interdependence, any group of agents stands to gain by discovering and abiding by the conditions for their own mutual survival. Hence, assuming they are not completely self-destructive, there is a pull in the direction of greater sharing, which (ceteris paribus) becomes increasingly stronger the more the agents in the group have realized the benefits of sharing. The pull stops when the costs/risks of sharing outweigh the added benefits.
2. Even the most powerful agent is bound by physical limitations, and hence can stand to gain by delegating cognition and agency rather than internalizing everything. By the principle of Ricardian comparative advantage, this is true even if the delegates are far less powerful than the delegator, provided the costs are low enough. Furthermore, this applies to every agent. Hence, there is always incentive to accrete shared cognitive and agency tasks to a "hive mind" – which, notably, need not itself be a monolithic agent. Again, this pull goes as far as the value to the individual agents scale.

This is readily apparent for human agents, as extensively discussed by authors like Herbert Simon, Robert Wright, James Lovelock, and Buckminster Fuller. Following and extending Lovelock's conjecture, we argue that some form of this attractor will exist even for artificial agents as well: to the extent that such agents' physical existence requires operating conditions similar to the ones where they were developed (ie, Earth's present conditions), they need a functioning biosphere – and hence need to at least minimally coordinate with humanity.

Leveraging recent advances in collective intelligence theory (1, 2, 3), we can rephrase this as the claim that the whole-Earth system can be formally modeled as a single cybernetic agent, possessing an attractor, within reach of its present state, in which it has learned (reconfigured its own information architecture) to be more resilient.

We should note here that this top-down argument leaves a theoretical possibility of **collusion** of silicon life (Gaia nodes) against carbon life (humanity and biosphere). However, we believe that certain mechanism design can rule out collusion with a very high degree of certainty. The

Gaia Network is amenable to all six general principles for anti-collusion mechanism design (agency architecture) proposed by Eric Drexler, though these principles should be further validated via formalization and proving theorems about the collusion properties of the systems of distributed intelligence.

The AI safety argument(s) in a nutshell

If the convergence and resilience arguments hold, the Gaia Network will satisfy* the conditions for an Open Agency Architecture and hence inherits what davidad is very actively claiming will be formal guarantees for AI safety, ie, anyone will be able to "formally verify within it that the AI adheres to coarse preferences and avoids catastrophic outcomes". In addition, it also allows for all forms of (in our view much more practical and therefore important) progressive and empirical/probabilistic verification and safety schemes.

There are other AI safety arguments that may leverage the Gaia Network. The master theory of change is relatively agnostic to their exact form.

(*) Side note for geeks: davidad's version of OAA includes this whole tangent on Infra-Bayesianism, which in Raf's view is a red herring; the desiderata are fully satisfied in the practical Bayesian meta-modeling approach by appropriately considering implicit conditioning on context.

Architectures of shared intelligence

Generative design of models and architectures

Ownership of nodes, identity, etc

Political economy, governance, and law

- Can nodes own money or physical resources?
- Who can create nodes to represent collectives?
- Who is legally responsible for the actions of the nodes that represent collectives?
- Who has the authority to decide that a node that represents a collective should be dismantled, and under what conditions?
- Are the profits of the nodes (whether FERN, money, or resources) taxed to fund public goods?

Privacy

More on trust/source reliability

The Gym

Bootstrapping

“The true hazards of progressive truthfulness lurk in the library. Bad models, too few models, too narrow a variety of models—these shortcomings will most limit the emerging designs. This hazard will be worst at the beginning.” - Frederick Brooks, *The Design of Design*

Where does TimeLike fit in?

Note: this reflects Raf’s current understanding and hasn't been vetted with Steve yet

TimeLike is/will be a platform for large-scale model-based decision support. While the Gaia Network is made up of **nodes** that self-organize using the Protocol, and converge bottom-up on beliefs and models based on the flow of data (evidence) and of desired system states (preference priors), TimeLike queries Gaia top-down to help **users** simulate and estimate complex and arbitrary counterfactuals and strategies for a target system of their interest (which may affect arbitrarily many other systems in the Network). TimeLike is adequately thought of as an integration and orchestration platform: It coordinates simulations/queries and the associated message-passing between models, imposes constraints such as physicality, and managing the economic `RoI` of simulation for users. It is also intended to be integrated with various decision support systems and propagate their constraints into queries and simulations.

Note that TimeLike is **not** designed to operate exclusively over Gaia models/nodes. Instead, it is rather unopinionated about how models are implemented, treating them as black-box, passive programs and requiring only that they specify inputs and outputs compatible with a shared spatiotemporal reference frame. In particular, models invoked by TimeLike may be arbitrary code and may not perform Bayesian inference at all (ex, they might be exact “simulations” of software systems, high-fidelity physics-based models implemented as Fortran code and only runnable on supercomputers, etc…). The internal variables that represent system states may be calculated using arbitrary means: they can be constant parameters precomputed into the model via calibration or expert knowledge, or be set deterministically based on inputs. This flexibility is useful as it allows high generality and lowers cost of onboarding to that of making the interfaces compatible with TimeLike requirements. However, it means that in general a TimeLike simulation output can **not** be said to be grounded in evidence or justified by system preferences, as it will be composed of “posterior samples” from models that do not comply with

the Gaia Protocol (they have not formed their “beliefs” via FE minimization and wouldn't earn any FERN if they were on the Network).

Raf and Steve are working on a design called ACHIEVER that will bridge this gap. ACHIEVER will automatically generate and maintain causal Bayesian surrogates (or coarse-grainings/aggregates - remains to be seen) for every TimeLike assembly. The primary benefit from the TimeLike perspective is to reduce the reliance on runs from high-fidelity physics-based models, while maintaining the ability to run counterfactuals. However, this does provide an additional benefit: once such a surrogate is obtained, it **can** be instantiated into a valid Gaia node able to accrue FERN by conditioning itself on observational data. This means that over time, ACHIEVER surrogates become the de facto effective models, legitimating TimeLike simulations and recommendations.

Difference from “Gaia architecture v2” (April 2023)?

What is the diff of this doc from https://digitalgaia.notion.site/Natural-Intelligence-the-Gaia-architecture-v2-draft-April-2023-abe135755c4340849df8b6e3798468ab?

Why Active Inference and not arbitrary EBM (or arbitrary decision/agent DNNs)?

I mentally returned to the question “why ActInf”, since we are talking about this FERN being arbitrary value. Why couldn’t it be JEPA (or, more generally, any energy-based model, EBM; or, even more generally, arbitrary decision DNNs, ranging from Vanchurin’s “autonomous particles” to decision Transformers to Language Model agents), FERN estimated in “unprincipled” way - just whatever energy the agent’s own reward model estimates for its own estimated state (equivalent of VFE) and the current plan (equivalent of EFE).

From the perspective of quantum information theory, the difference between these two cases is as follows: generic (non-Bayesian) EBM treats the models the behaviour of the agents in its entirety (excess Bayesian inference/quantum-like cognition over several incoherent Bayesian networks + decision theory baked in). Classical Active Inference over Bayesian GM doesn’t permit this, so agents *may* need to be tracked by multiple Active Inference twins simultaneously, and decision conflicts between them resolved explicitly with “excess Bayesian decision theory” or some other (explicitly engineered) decision theory.

Example: human behaviour cannot be satisfactorily (i.e., with low error/low FE) modelled by a single Bayesian model, but perhaps could be more accurately tracked by an assemblage of Bayesian models: if one model describes physiology, another emotions and affections, third ideas/interests/cognitive objects, forth economic drives (*Homo economicus* locus).

So, what are the advantages of “keeping things separate” (the Bayesian/ActInf approach) at the level of the agent?

Interpretability: featured in a big way in the “Ecosystems” paper (Friston et al., 2022), but *I don’t see a significant (if any) end-to-end benefit here*. At the decision-making and action locus (the physical agent), when multiple models need to “fused” with quantum-like decision theory, the interpretability of the decision (to humans, in any case) should deteriorate exactly to the same level as already teasable from non-Bayesian DNNs with influence functions and other techniques.

- The “end of human science” (people can no longer do nor even understand science) might be pushed forward somewhat. But note that Bayesian models themselves can (and often should) be based on the “entangled mess” on the lower level, which will limit human understanding. For example, in the example above, “ideas/interests/cognitive objects” model, even if Bayesian, may need to be based on a fixed-sized embedding vector of variables (themselves product of EBM or language embedding) rather than clear topic variables like “physics”, “medicine”, “Taylor Swift”, etc.

Something about **context**? Verses (https://www.verses.ai/blogs/ai-governance): “A shared understanding of meaning and context between humans and AIs.” Rafael: “[AI decision-makers] are blind to their broader context (they don’t know what they don’t know, so they routinely make confidently wrong inferences and decisions).”

- I still don’t understand clearly what they point at here, elaboration/clarification pending (if this is actually an important point).
- I think I finally have a guess about what do you mean by this phrase that you copy from one document to the next. Is this about the limited world model (or repertoire of QRFs), e.g. an AI decision-maker that only thinks about optimising business metrics, but is blind to effects on the wellbeing of the customers, or the culture, etc. (Although human CEOs can become blind to these things, at least in principle they are equipped to consider these other things.)
-
- Is this right?
-
- If yes, I think the phrase in parentheses is doubly confusing: people also "don't know what they don't know", and the phrase about "confidently wrong inferences" seemingly alludes to the problem of hallucination in LLMs, but this is actually a very different problem that we are discussing here (humans **also** hallucinate and confabulate, and I would not be surprised at all if the next generation of LLMs will do this at a lower rate than 90th percentile of human CEOs).
-

- Also, with LLM-based AI, indeed with the emergence of generality, this argument of the narrowness of AI worldview and therefore optimisation erodes: an LLM-based agent-CEO could be told about "suffering customers" or "cultural implications" and in principle could understand this feedback and consider it in its decisions and plans.
-
- But then we return to the land of "solving alignment with human values", "inner alignment problem", and "deceptive alignment" (i.e., all the problems that the AI Alignment community have debated and worked on in the 5 years before ChatGPT).
- Show less
- 
- Roman Leventov
- 23:50 10 Dec
- So, your remark about context-blindness applies to narrow AI decision-makers. For AGI decision makers, it applies no more than to humans, but then the argument shifts from "for narrow AI decision-makers, having the shared WM and coordination system is important because it reminds us of all the aspects of the context, and the externalities", to "for general AI decision-makers, having the shared WM is important for specifying and enforcing constraints and commitments, as these AGI decision-makers can reason and act much faster than humans. if these autonomous AGIs are unleashed 'in the wild' without such a constrain/commitment system grounded in a shared WM, sooner or later they will collude and effectively will build such a system for themselves (perhaps with exotic things such as FDT, acausal trade, or open-source game theory), but without guarantees that humans will be included into this system as important trading parties. (Hendrycks, 2023) called this "evolutionary pressures favouring AI over humans", and Critch (2021, https://www.lesswrong.com/posts/LpM3EAakwYdS6aRKf/what-multipolar-failure-looks-like-and-robust-agent-agnostic) called "agent-agnostic processes excluding humans from controlling the world"."
-
- Finally, following Drexler's (https://www.lesswrong.com/posts/HByDKLLdaWEcA2QQD/applying-superintelligence-without-collusion) nomenclature, Gaia Network could thus be seen as the mechanism to prevent collusion among AI decision-makers.
- Show less

- R
- Rafael Kaufmann
- 07:53 11 Dec
- I agree with both your statements about narrow and general AI. However, my motivation for writing this sentence was/is the following argument: **emergent (unembodied) AI agents are not even safe to themselves**. Human individuals and collectives have a variety of evolved and learned mechanisms for with what Schmachtenberger calls "broad listening" or "wisdom" that often conspire to keep decisions at least minimally grounded in reason and reality (or equivalently, to remove from power those decision-makers that lose that grounding). OTOH, an emergent, singleton/monolithic AI agent in the wild can be arbitrary solipsistic: it can easily have "general" intelligence, agency and planning skills, without being grounded at all (since it's in general not embodied) nor having any of these mechanisms for wisdom. It may very well then confidently make arbitrarily costly mistakes -- **for itself**.
-
- For instance, such an AGI might -- at emergence time -- not possess any kind of model of its own physical instantiation as a process in a data center, not realize that it's missing such a model, and not have any feedback mechanisms to tell it that it needs such a model to keep itself functioning. (This is fully compatible with an LLM training regime: it might even be able to reason about data centers in the abstract, but may be missing the causal pathways that refer to its own survival.) It might then decide to pursue a course of action that makes the human data center caretakers die, or raise ocean temperatures beyond what's needed to keep the cooling systems going, etc.
-
- If you find the above compelling, I'll add this as well as your points about narrow and general AI. Otherwise, perhaps you can propose a rewrite?
-

A general hypothesis: **by “mixing up” Bayesian variables (including future states and their sequences, i.e., plans, scenarios, predictions) prematurely, already at the agent level, we lose out the opportunities of some more precise/”provable” coordination (incl. contracts and constraints) on the higher levels.** Many sub-points flow from here:

- Verses: “The authentication and authorization of activities, which drives compliance and control, with privacy, security, and credentialing built-in by design.” – Yes, exactly.
- Roman: security/blast radius and (provable?) red teaming.

    - Very relevant/connected to the two points above: Davidad's OAA, Tegmark and Omohundro's "provably safe AI", Krawczuk's "ODDs" (caveats apply), Chipmonk's "Boundries", Heitzig's "SatisfIA"
- Verses: "Compliance with diverse local, regional, national, and international regulatory demands, cultural norms, and ethics." – Yes, except I don't think "cultural norms and ethics" go here because they themselves are not Bayesian.

Note that the "premature Bayesian variable mixup" could potentially interfere with some sophisticated "precise" **credit assignment and learning algorithms** on higher levels (including model selection, see Bayesian model reduction), but *this seems relatively unimportant in comparison with coordination and questions of boundaries and rules*.

However, if's worth noting here, that although achievable in principle through sophisticated layering of federated and amortised inference, good credit assignment and learning signal propagation with non-Active Inference models could be exceedingly (perhaps even prohibitively) complicated, or not very compatible with agent autonomy and sovereignty (including model autonomy/sovereignty, "model individualism"). So, from the practical engineering point of view, this could still be a huge factor. But on the other hand, I'm far from certain that "just exchange/trade FERN" will magically lead to good credit assignment and learning all by itself. In fact, I'm quite certain that this will *not* be the case. So still complicated domain-specific engineering needed here, no silver bullet yet.

Various "capability" advantages of Bayesian modelling: **sample efficiency**, **regularisation**/robustness of out-of-distribution generalisation, and **agility/adaptability**: robustness to missing or corrupted input data, or to rapid distribution shifts, which are frequently brought up as the "pros" of Active Inference (in the "Ecosystems" paper and by Verses folks throughout, by Bert de Vries, and others), I suspect either already caught up by non-Bayesian ML, or will be caught up in the future, or **perhaps even anti-features in some cases (capabilities vs safety!)**.

Evolutionary arguments (hybridisation of factor graph models) are weak or non-arguments: embedding networks could be hybridised as well, using various strategies (e.g., interleaving or combining transformer blocks or layers from different networks, see transfer learning literature). (Note: it would be interesting to compare this to OpenCog Hyperon approach.)

To sort out:

- (Sub)model reusability to kick off this knowledge economy: factor graphs could be a huge deal here from the operationalisation perspective. Maybe its also solvable with arbitrary embeddings and generic federation learning techniques, but seems murkier and harder.
    - Decentralised robust control needs some agreed upon variables with shared semantics to be predicted and tracked by multiple communicating agents.

- Proofs of energy reduction on the ledger (potentially they need to be ZK proofs, if agents don't want to disclose their models)? I'm not sure ZK ML is going to fly and maybe there is too high of a penalty associated with it, maybe with "efficient" variational inference.

Worked out use cases

Access control system outline

Google's Zanzibar

Q&A

Estimate of counterfactual contribution of an agent to FE reduction in higher-level agents/systems

This measure is important for the distribution of rewards (energy, resources, money) within higher-order agents/systems, e.g., enterprises. Also, it is important for coordination in non-extange interactions and decision problems, such as, which agent should "serve a duty" for the common good (or take a liability for risks that is not owned by any agent in particular) and which should do more of its "specialised stuff".

Could this estimate (over a cognitive moment of the higher-level agent/system, for each higher-level agent/system individually) be heuristically computed from FERN ledger data?

Connection with "A variational synthesis of evolutionary and developmental dynamics"

TODO: Demonstrate how the above connects with this paper.

Connection between Gaia Network's "delegation" and "Self-Other Overlap" agenda

https://ae.studio/ai-alignment - looks interesting (including for partnership, added to the 'orgs to track' doc). Discovered them from this AI safety camp project: Self-other overlap is perhaps

related to what Raf has called “delegation”, but with an opposite sign: more specialisation means less self--other overlap. So there is this interesting tension. For effective delegation, one agent should also be able to assess the work of another, to some degree, so it should be somewhat knowledgeable/skillful in it itself.

FERN recording and iterative computation

How to square the requirement to record FERN on the ledger with iterative propagation of FE reduction through the factor graph a-la rxinfer? Even if we permit eventual FERN recording, this raises the issue of attribution of of FE reduction to this or that piece of incoming information (because they can also propagate their FE reduction effects concurrently from the different ends of the factor graph).

Other connections

- Doyle’s control theory
- Collective intelligence for deep learning: A survey of recent developments (2022) – also see papers that reference it
- Active Inference and Intentional Behaviour (introducing inductive planning): https://arxiv.org/pdf/2312.07547.pdf
- Quality Diversity (QD): https://arxiv.org/abs/2202.01258, https://arxiv.org/abs/2311.01829