On the simulation pipeline and its consequences

(One of two essays towards correctly trusting AI)

There's a lot to the problem of correctly trusting AI. Knowing "how AI thinks" or whether it does isn't our primary concern, rather we should seek to understand how it processes information and the effects it produces. The distinction between thinking or not is of great importance, *in situ,* that is to say all occurs in a situation.The view from nowhere is a view of somewhere after all - it's a matter of framing, what Bree Beal would call a *polyphonic* situation.(Breal 2022) Here we must frame around models and heuristics, the former helps us see what is there, the latter helps us use what is there efficiently, but both are procrustean,downstream from bounded rationality - there is always an excess. So *in situ* becomes the name for the set that contains *in vivo*, *in vitro*, and most importantly for this

essay, *in silico* and *in stilo*,(in silicon and in writing). These two frames are derivative from the distinction now considered archaic between digital and human computers.A human computer doesn't translate words in numerical chains all the time, nor is it their only only option, but there will be moments of similarity between a human and digital computer but the significance of the meaning of symbols like the 'red, white and blue flag' immediate evokes ideas like 'not mine' and/or 'ours' as the french, russian and american flags are not all mine, whilst the British is ours. *In stilo,* I can depict my sentiments more easily with the word 'ours' (to a noncoder) than an *in silico* depiction of vectors and weightings. *In silico,* more effort is needed and the question remains is my 'feeling' for that flag best in all situ depicted in - as yet - silicon incompatible terms? Will the particular 'sentiment-without-numerical-translation' terms be incompatible indefinitely? I am inclined to say to a certain extent mostly yes.Framing/polyphony - in other words - context - is indefinitely required.

The correctly trustable AI will be downstream from its construction as able to function wisely, or those engaging with it know how to be in a quasi-Aristotlean phrasing - technically and/or theoretically wise. The distinction between rationality as *instrumental* and *epistemic* is part of the reason why automatic entities will not just solve the problems or place us indefinitely within the basin of good governance as people are persistently irrational, depending on what we mean by rational. Correctly trusting AI as standalone either requires acknowledgment of the historical or contemporary context. To deal with the homophone problem, for example as found for example in virtue ethics for example, bravery is a set that contains vicious and virtuous forms: bravery, cowardice, and arrogance are all forms of bravery with recklessness, arrogance and cowardice as extremes. Similarly intelligence can be virtuous or vicious, with analysis paralysis being one extreme and hubris as another. Only it's not so simple: the extent to which one can bes virtuous by performing a virtuous action and still not have a generally virtuous disposition complicates matters, especially if Sukaina Hirji is correct to say that what Aristotle's Nicomachean Ethics do not contain is "a distinct normative theory alongside deontology and consequentialism."(Hirji 2018) Thus we must ask if it is wise to efface the distinction between *in silico* and *in vivo*. Imagine a digital model is produced of a particular human body, a surgery is run in simulation, has that surgery been performed *ex silico* by that fact alone? No. If a surgeon performed a different surgery simultaneously as the *in silico* surgery would the same thing have been done? The phenomenon it aims to capture is not the phenomenon itself.How ought we trust the shadow surgeon in Plato's cave?

What is AI?

For Herbert Simon, AI, is a simulation, an artifactual depiction at best. Even integrated into a system, as steerer, there is a difference between creating a program of activities and responding to the immediacy of external activities. Thus a cybernetic system whilst it may utilise AI, needn't be AI or artificially intelligent in the idiomatic manner we are accustomed to.

How can we trust it?

I agree with Ramón Alvarado that trusting AI on the basis of a quantification of probabilities of benefits, is a suspect argumentative strategy "as is well known in ethics appealing to 'lives saved' simpliciter could be used to justify an uber-paternalistic tyrannic form of government, some forms of slavery, or other highly undesirable forms of social contract."(Alvarado 2022) We must take framing into consideration, even if according to the polyphony principle the potential to frame an event differently, precedes any actual framing.Simply using quantitative or qualitative formal logic to frame a situation isn't going to prevent disasters.(Beal 2022) For more on the difficulties of framing one might consider Bermundez's three hypotheses::

(H1)Framing highlights one key aspect, guiding a specific response and affecting decisions.

(H2) Different frames provoke different responses, creating conflict that can't be resolved easily.

(H3) In complex situations, framing effects and cyclical preferences make sense in situations

(Bermúdez 2022)

Trusting AI is not as simple as trust in prediction. Rather I agree with Alvarado that AI is an *epistemic enhancer* and that

1. It is a technical artefact that is designed to expand our epistemic capacities in at least two ways: it is deployed in an epistemic context such as inquiry; and the capacity that it expands is the capacity to calculate or compute and not dig, mix, or other physical tasks.

2. Content: AI deals with epistemic content: propositions,models, vs. mere symbols variables/numeric values.

3. AI is also the kind of technology that carries out epistemic operations, such as analysis, prediction, and inference. (Alvarado 2022)

The type of thing that AI is, is relevant to how we ought to trust it. We wouldn't trust a dog to do battle with a sperm whale. To define what we mean by "correctly trust" I might suggest Alvarado's position: 'we trust it correctly when we trust it in virtue of what it is.' If AI is "a data analyzer pattern recognizer or inference and decision-maker" as Alvarado suggests "a high-level of trust in the results of computer simulations is only appropriate when they are grounded in well-curated data, plausible scientific theory, empirical evidence, and good engineering practice. Though these normative criteria may seem straightforward, they are difficult to follow in scientific practice and they have been underemphasized in philosophical discussions of the epistemology of simulation. Simulations that are not grounded in these ways, may turn out to be interesting sources of insight or inspiration, but they **should not be accepted as guides in decision-making where potentially harmful or expensive risks are involved."** (Symons & Alvarado 2019, emphasis added)

Here we should consider the point that Ernst Friedrich Schumacher said in *On Technology for a Democratic society* “If you want to be a good shoemaker, it is not good enough to make good shoes and to know all about making good shoes, you also have to know a Jot about feet. Because the aim of the shoe is to fit the foot..” (Schumacher 1982) What can AI know about having feet? It can run the numbers, but there will be a non-trivial gap between the numbers and what it's like for a shoe to fit. The fact that our next word choice mechanism isn't completely reducible to quantitative weights means that we have no immediate rational justification to not align with Alvarado when he says “For example, policies regarding large-scale institutional interventions, life-critical systems assessments,existential risks, and the outcome of otherwise untested medical procedures would fall within [the] category” of being a source of inspiration and insight but not an empirically or rationally rigorous source of knowledge.(Symons & Alvarado 2019)

If we absolutely must trust AI,we should understand, ontologically, what type of instrument it is. Understanding how AI generates output—beyond its numerical translation capability—raises the question of whether the way it processes and expresses words differs significantly from our understanding. Can a machine have an emotional response to the emergence of new perspectives? How can the affective aspect of “the good life” be considered by AI and how will the advice change the more the desired outcome is of an emotional nature. Now one may say that automated wisdom and philosophy don’t have as their output ways to live the good life, its merely a way of processing and analysing easily numericalised data, but then the advice or output we trust is limited to quantitative analysis, as humans do no live in a purely quantitative mode.Even a spreadsheet of purchases refers to more than sales and purchases. What does a shoe-making machine know of comfortable fit? The measurements of height, width, breadth and so on, it has a model but the wearer is not the model and the model may approximate the wearer but it wont have all of its experience, this is a non-trivial difference. Although RLHF may mitigate some of the issues, the nature of the simulation pipeline (*in vivo* to *in silico)* means that completely closing the loop, may be impossible or unwise in a great deal if not all contexts.

The simulation pipeline

If we accept artificial intelligence as automated digital calculation, complex information process(or) or a simulation, a question arises of: what is a simulation?

For Nicole Hartman: “Simulations are closely related to dynamic models. More concretely, a simulation results when the equations of the underlying dynamic model are solved. This model is designed to imitate the time-evolution of a real system. To put it another way, a simulation imitates one process by another process. In this definition, the term “process” refers solely to some object or system whose state changes in time. If the simulation is run on a computer, it is called a computer simulation.” (Hartmann 1996) For Christopher Hubig and Andreas Kaminski simulation is “is the aggregate of transfers of elements and their relation from one representation in another for the purpose of expansion, revision, and alteration of our theoretical and practical references to the world.”(Hubig and Kaminski 2017) For Paul Humphreys something is a simulation when it is a system that “produces, via a temporal process, solutions to a computational model”, that correctly represents a real process or object either dynamically or statically.(Humphreys 2004) Whilst Nicole Saam sees simulations as having two forms, one which resembles the epistemology and methodology of thought experiments and the other the epistemology and methodology of material experiments (Saam 2017). What all agree on is that *the simulation model is not the simulation process and neither are the results,* regardless of the attractiveness of Stafford

Beer's dictum (the purpose of a system is what it does),(Beer 2002) one should remember that nothing does the same thing in all frames. To Beer's dictum we should add the postphenomenological notion of *multistability,*(Wellner 2020) that any thing is for more than one thing, thus whilst a thing produced with planned obsolescence is a money making scheme, it is both a labour saving device and (an eventual) source of scrap.

What a simulation models and what model a simulation uses are distinct elements in what is called in philosophy of science the 'simulation pipeline.' For Michael Resch the pipeline is reversible and goes

1. Reality
2. Physical model
3. Mathematical modelling
4. numerical scheme
5. program structure
6. programming modelling
7. hardware architecture

For Eric Winsberg:

1. Theory
2. Model
3. Treatment
4. Solver
5. Results

For Nick Szabo:

1. Physical reality
2. Mathematical model
3. Numerical solution
4. Prediction

Resch says that each of the three models above "when compared with the other two exhibits considerable gaps… However, upon closer examination…a homogenous concept of simulation does not exist." (Resch 2017) If it's axiomatic that AI is a simulation or the procedure for producing one the notion of conducting theory-free experiments or treating

simulation results as epistemically equivalent to experimental results poses significant dangers. Even if we assume the simulation's model is entirely free of theoretical bias, what we do with its output is guided by intentions rooted in heuristics, biases, and underlying models. This raises questions about the expertise of those utilising the output and the significance of refining the model. Whilst common sense theories (i.e.,models) of Gramsci, Foucault and Heidegger, may not be of great interest to AI engineers, we're being asked how best to represent them in code. The utility of Bayesian statistics notwithstanding, the probability of something like disgust, isn't identical to the mechanics of its existence across the species, nor the particular way in which disgusting is a salient element of an object, for one who feels it. To build trustable AI requires not ignoring the rationalist for the irrationalist or vice versa but for respecting the virtue and limitations of varying orders of operation.The irrationalist who creates a model based on rationalist work and vice versa, seems more likely to accurately and pragmatically model the experience than those built in either silo.

In trusting AI, we must distinguish between the simulation and the simulation as steering a system. Artificial intelligence and social cybernetics must at times be considered two linked but distinct instrumental and epistemic domains:. The know-how vs how and the know-that.This is somewhat analogous to the difference between a system of feedback and one of reciprocal activities. Ursula Franklin highlights the difference saying that "reciprocity is not feedback. Feedback is a particular technique of systems adjustment. It is designed to improve a specific performance…the purpose of feedback is to make the thing work. Feedback normally exists within a given design. It can improve the performance but it cannot alter its thrust or the design. Reciprocity, on the other hand, is situationally based. It's a response to a given situation. It is neither designed into the system nor is it predictable. Reciprocal responses may indeed alter initial assumptions… lead to negotiations…and they may result in new and unforeseen developments." (Franklin 1999) To say that a cybernetic system is reciprocating, in the same way, people in conversation are is to take Paul Grice's notion of conversational implicature as ontological - as rules rather than tendencies and guidelines.The difference is the degree.

Assuming that the concept of a simulation pipeline accurately captures reality the issue of transfer effects becomes significant. If the theory doesn't perfectly match the model (qualitative, mathematical or digital) then multiple translations/transfers are involved. Hubig and Kaminiski (H&K) are correct to claim that "the question then arises of how to ensure the transfers (2-7) simulate the target system."(Hubig and Kaminski 2017) H&K have simulation model as well which they present in a simple and precise form: step 1 is "the physical-qualitative model" that "simulates a segment of reality"; step 2: "the mathematical modelling" where "the physical model is simulated in mathematical model"; whilst step 7: "forecast of facts and situations that culminate in appropriate actions plans" which "forecasts simulate the expected systems states."(Hubig and Kaminski 2017) The point at which an *in silico* or cogito model becomes *in silico* could be placed at step 3, but what's important is that at each position a transfer takes place. For H&K "the justification for when a sufficiently large and sufficiently certain correspondence between each simulation and the simulated system/model is present cannot be dealt with without considerations of a theory of truth." Because simulations are instruments (not theories or experiments per se) the justification of a simulation for H&K is a matter of adequacy. Problems with parameterisation occur not only due to translation effects and artefacts, but because "simulation models are adjusted to empirical data…the danger of overfitting arises for unknown (unassessed) areas" . In cases of simulation, we contaminate our epistemic rationality by utilising it for the discovery of truths - especially for an instrumental purpose. This is not to say rationalist projects have no boon to give, but as a commitment, insofar as truth is out there, there is no *a priori* model for finding all of them at all scales that reduce to the true as easily translated into the numerical. Part of solving (meta)philosophical problems, well enough, requires a project of translating James Ladyman and Don Ross' scale relative ontology into domains beyond philosophy of science (Ladyman et al. 2007). If more people are comfortable and willing to articulate the

scale, scope, domain and preferred framings of their analyses, we'll spend less time arguing and more time finding out if and how models can fit together. I doubt the utility of anything like Leibniz's idea whereby formalisation of everything can be completed and results in every object being given a multi-domain coherent numerical form. In so far as no natural language has been found that lacks polysemy, I am not convinced that a total monosemic language could allow us to communicate well.

Part of H&K's justification for a pragmatic theory of truth and error in considerations of simulation is that: "The question of when a chain of transfers is sufficiently consistent and coherent…cannot be answered in formally logical terms, but only in practical terms as soon as approximations are involved." To take a pragmatist approach to truth is to question the honest ability to virtuously fulfil the desire to be solely epistemically rational with regard to epistemic commitments. What needs to be remembered is that in making trustworthy AI, we must account for the fact that moving into *in silico will not be done without* shadow or errancy. It seems that no epistemic result or conclusion occurs without a context of instrumental means and ends. I see no reason, other than out of a desire for incestuous purity, for one to consider the rationalist brand of epistemic rationality as the only good one, and that rationalist alignment is significantly less likely to be enmeshed with a pathological bias than any other. The problem of accounting for transfers is not solvable only by rationalist means. Translating epistemic oughts to instrumental strategy may involve attending to plurality of framings earlier in the pipeline than is currently done, if it is at all.. As Johannes Lenhard puts it: "The success of simulation modelling hinges on iterated adjustments instead of mathematical derivation, and it proceeds by pragmatic amendments (parameterizations) rather than finding 'the right' mathematical structure. … analytical transparency is seriously questioned by the very methodology of simulation, since iterated feedback loops during the modelling process make it hard to attribute particular behaviour to particular assumptions. …the advantage of the computer,... simulation modelling over traditional mathematical modelling, is based on the speed…this does not simply extend the conception of mathematical modelling, but re-structures it in fundamental ways" Though not intended as vindication of Hubert Dreyfus,Lenhard's point,aligns with Hubert's view that targets should not be treated as identical to the phenomena they represent, especially when modelling.

The changes and artefacts that result from transfer and translation stem from discretisation techniques - which Symons and Alvarado say include "epistemically relevant decisions on the part of the modeller that are distinct from the original mathematical model." But these decisions and their constraints are not the same as the computational on how a model is digitised. At the level of epistemic-instrumental rationality this implies that Alan Turing says "we cannot so easily convince ourselves of the absence of complete laws of behaviour as of complete rules of conduct. The only way we know of for finding such laws is scientific observation, and we certainly know of no circumstances under which we could say, 'We have searched enough. There are no such laws."

And when Yudkowsy says "Bayesian formalisms in their full form are computationally intractable on most real-world problems. No one can actually calculate and obey the math, any more than you can predict the stock market by calculating the movements of quarks." Like when Simon says of bounded rationality that eventually "we will begin to interpret as rational and reasonable many facets of human behaviour we now explain in terms of affect." Each is not necessarily a proof of a mathesis of human intentionality as it is a praxis being articulated in purely formal logical and numerical terms. For Simon, a role consists of decision premises not the decisions made from those premises. This means that when Yudkowsy expects Bayesian-style belief updating to achieve closer correspondence between map and territory, we must ask whether k-complexity is the best way to measure the efficiency and effects. Also, are these measures generally efficient and/or desirable for all aims and contexts? Dispersed knowledge is valuable, but dispersion doesn't guarantee

quality. The utility of rationalism, with its tendency towards formal symbolic logic and numericalisation doesn't mean that the base claims of it are epistemically warranted by virtue of their utility. Hence why John Symons and Alvarado make a point to distinguish between *in silico* simulations "grounded in well-curated data, plausible scientific theory, empirical evidence, and good engineering practice" and others, and why Björn Schembera and Juan Duran separately address the limitations of simulations for explaining all the elements in social segregation in their chapters of *The Science and Art of Simulation.* We can make a distinction between the natural sciences (naturwissenschaft/ natural philosophy) and humanities (geisteswissenschaft/ practical philosophy) or simply say that there are fields and studies that are mostly quantitative or mostly qualitative. Regardless, the instrumental virtue of a semi-qualitative but rational strategy - if possible (see continentals have a sense of humour) is context specific, and to the extent that humans as bodies as lived by subjects are considered to be more of that than as brute body objects, the purely quantitative rational approach is not enough to steer us into a basin of good governance, but it must play a part.

If the rationalists argue that numbers and symbols are the blocks of truth and the phenomenologists say concepts (all, some, being, and italicised greek words) are the blocks of truth, if only for the sake of peace,we should say the bottom level is of numbers and letters.The mathesis sought by Leibniz must have formal and informal parts otherwise the distinction between regularity and elements of surprise become meaningless, and we are Aristotle's theoretically wise, knowledgeable but not useful.

This suggests that virtuous epistemic rationality has a non zero chance of occurring from a combination of mathematical and non-mathematical means, and as such the virtue of rationalist analysis does not dis-warrant ir-rationalist analysis, even if, *in silico* models and simulation need formal logic. The rationalist, the developer, artist, politician etc if they are roles with decision premises, also possess unique languages and lexicons.As Henrik Sinding-Larsen put it : "To a greater extent than ever before, language has become a question of conscious choice: we can evaluate the properties of one language against another. Through …the construction of programming languages, language has become an object of construction and invention." (Sinding-Larsen, 1991) Thus, an epistemic warrant may be derived from any number of frameworks. Each stage of a simulation pipeline has independent epistemic (and instrumental) warrants . Symons and Alvarado say "Even if the mathematical model is fully warranted and works as intended and even if the discretized version of the model also works as intended there is no reason to think that we have reason to trust the latter because of the former. When a discretized model is ultimately implemented in a device, for example, it requires yet another epistemically relevant transformation in the process of coding. In coding, considerations of fit, trust and/or reliability of a given algorithm will depend on independent factors from those involved in the discretization process. Unlike discretization techniques that involve established techniques and theories, code is often the result of highly idiosyncratic problem-solving approaches."(2019,14)Neither a simulation nor its output inherit credibility through association but each step must be justified, with their detriments made clear. This principle applies to the pipeline of mind to pen to paper to conference to policy to society. No-thing is always (already) perfect.

In correctly building correctly trustable (wise) AI, I think the creators of *in stilo/silico* models and their subsequent should keep in mind something like Mel Andrew's model of model transfer as presented in *The math is not the territory* , the parts of which are: structure-equations or axioms; construal – what a structure is used for; and reification – when an analogical relationship is taken literally or elements of construal get put in another domain. Mel Andrews argument should give pause as to the notion of producing wise AI.Beyond simulating the absence of mistakes or surpassing non-experts, we should ask is the foundation of a simulation (as code) fundamentally different from its execution and output?Asserting their isomorphism extends the computational theory of mind to an absurd

conclusion: that we have already created something that is consistently equivalent to human thinkers across all domains and contexts.

Whilst I don't agree that AI moves us away from the vector of a good life, I agree with James Nguyen and Roman Frigg that "while a structure that's mapped to a target is not ipso facto an explanation, the existence of such mapping is a precondition for (at least some kinds of) explanation."(Nguyen & Frigg 2017,19) A computer's predictive utility doesn't justify taking it as isomorphic with the mapped territory. For example, we accept that "[c]arbon, hydrogen and covalent bonds are bona fide natural kinds." and this is a formalisation of a methane molecule, but this may produce non-isomorphic structures. Many structures can map to a target, thus if a map is not the territory, nor is the model for the map the map (e.g., the way a 3D structure is represented in 2D Peterson and Mercer maps).

To extend this idea to human thought experiments as a form of simulation, we accept that human and computer simulations operate in the same domain. This domain requires parameters that distinguish thought experiments "from other real patterns in the same domain." Both human and computer simulations are in the same domain: derived from models, although they can be distinguished from each other based on frame and scale. Exploring scale-relative ontology is a larger discussion and would require more expertise (as suggested by Mott). An idea worth considering might be whether a model represents a pattern and to what extent can multiple models be useful and valid interpretations of the same pattern.

On Rationality

Creating "wise AI" may first require a bit of dirty work, what Mary Migdley called *philosophical plumbing.* "Conceptual schemes as such are philosophy's concern, and these schemes do constantly go wrong." The philosopher as a plumber is not just a one time visit, however. They are to be concerned with ideas treated not as "stagnant ponds" but as "streams that are fed from out everyday thinking, are altered by the learned, and eventually flow back into it and influence our lives."(Midgley 2000) It's not only the pointing out of problems that Midgley considers the activity of philosophers but they are beholden "to notice what one is not noticing." Consider the Gramscian notion of organic and traditional intellectuals, philosophical plumbers may emerge through companies and forums like Less Wrong, whereas traditional plumbers may be found in academic contexts, but they may have an overlapping role.The prescriptive aim is to present and construct schemes that mitigate the detriment of pre-existing schemes. Midgley's conceptual schemes can be further distinguished into models and heuristics, both of which can have their holding justified by epistemically rigorous or affectively significant reasons, it is not my intention to eternally condemn either to intellectual irrelevance, but we should consider why we hold the models that we do, but more importantly the implications of not deviating from them.

We must be careful to not take an implemented model or heuristic or model to be implemented as sound or valid merely by its use or its success so far, especially without consideration for the relevant externalities, compare this by Ursula Franklin: "Production models are perceived and constructed without links into a larger context. This allows the use of a particular model in a variety of situations. At the same time such an approach discounts and disregards all effects arising from the impact of the production activity on its surroundings. Such externalities are considered irrelevant to the activity itself and are therefore the business of someone else. Think of a work situation, a production line. There are important factors—such as pollution or the physical and mental health of the workers —which in the production model are considered other people's problems. They are externalities"(Franklin 1999) With this by Milton Friedman: "In [a free] economy, there is one and only one social responsibility of business- to use its resources and engage in activities

designed to increase its profits so long as it stays within the rules of the game, which is to say, engages in open and free competition, without deception or fraud." (Friedman 2002)

This juxtaposition of Freidman and Franklin demonstrates how the construction of boundaries can be used to disavow effects resulting from one's action. If everything but deception and fraud is acceptable, then do we in general have no good reason to be concerned with the activity of businesses?

The juxtaposition also shows the accuracy of Simon's concept of bounded rationality: the idea that there are internal, intrinsic, and external limitations to the cognitive capacities of an agent. What happens when a heuristic that is rationally held, but like many if not all heuristics, is only contingently rational, and is therefore not multidomainoptimal?There is danger that a heuristic taken out of context of its model could be considered a universal. A prescription becomes a description, a heuristic becomes a model, and we lose the ability to consider, in a trade with permission to act. Of course, everything is somewhat both, but the point remains that a model is not a heuristic. This is important as the model of bounded rationality as the human way, can be expanded upon in different ways, one of these is Gerd Gigerenzer's notion we use our bounded rationality as heuristics or "a mental process that ignores part of the available information and does not optimize, meaning that it does not involve the computation of maximum or minimum. Relying on heuristics in place of optimising is called satisficing." (Gigerenzer 2010) That we use heuristics is not sufficient justification for the soundness of a given heuristic. Nor can we say the same for a model, asSimon says "the first consequence of the principle of bounded rationality is that the intended rationality of an actor requires him to construct a simplified model of the real situation in order to deal with it. He behaves rationally with respect to this model, and such behaviour is not even approximately optimal with respect to the real world." We must remember there are layers of bounded rationality in our psychology, our sociology, and our models.

Rationality is not total, and therefore we must be careful to not take an implemented model or heuristic or model to be implemented as sound or valid merely by its use or its success so far, especially without consideration for the relevant externalities.

This is why Michael Huemer distinguishes between Bryan Caplan's ideas on "Instrumental rationality (or "means-ends rationality") and "Epistemic Irrationality." The former "consists in choosing the correct means to attain one's actual goals, given one's actual beliefs." The latter, "in forming beliefs in truth-conducive ways —accepting beliefs that are well-supported by evidence, avoiding logical fallacies, avoiding contradictions, revising one's beliefs in the light of new evidence against them…." These two forms of rationality are necessary for Huermer's formulation of his theory of Rational Irrationality which "holds that people often choose rationality to adopt irrational beliefs because the costs of rational beliefs exceed their benefits" and that "it is often instrumentally rational to be epistemically irrational." Herbert Simon's hope with the model of bounded rationality is that more of the motivators categorised under irrationality such as emotion and affect will be understood under rationality. In *One-Dimensional Man* Herbert Marcuse says of advanced industrial society that the "irrational element in its rationality" is descriptive of many if not all trends of advanced industrial society. Philosopher of science Paul Feyerabend, and subsequent philosophers of science have also maintained that society is moving in this direction, with some AI-involved epistemologists taking note.

Whether or not one considers the profit or incentive or productivity imperative to in all situations, contexts, and amounts good, is irrelevant if one holds as axiomatic that within multiple framings or models, rationality in action contains or produces irrationality. The first thing that needs to be taken into account when considering trusting AI is that we, at our most charitable, will act in a somewhat less than maximally rational way.. Whether or not we are

‘in the loop’ with a wise AI is irrelevant because what it can do is the result of actions performed by persons who have, at the very least, a tendency to be irrationally rational. We should be more willing to interrogate and compromise on our heuristics and models.,If it is not feasible to make every cook a governor, we may do well to make every intellectual and practitioner capable of doing their own plumbing.

Even if we have a wise AI, the model it uses and the model used by the interpreters of what its output oughts are still entities that may hold irrational epistemic beliefs, thus prior to considering how to produce AI that can locate problematic ontologies and/or mitigate and/or resolve them, one should encourage humans to do so..To do this, we should consider that there are multiple forms of rationality, for example, Max Weber’s notions of

> *value rationality:* the determination of “a conscious belief in the value for its own sake of some ethical, aesthetic, religious, or other forms of behaviour, independently of its prospects of success”, and his
>
> *instrumental rationale*: the determination “by expectations as to the behaviour of objects in the environment and of other human beings; these expectations are used as “conditions” or “means” for the attainment of the actor’s own rationally pursued and calculated ends.”

These are two of four types of social action, the other two although considered by Weber to be at times “on the road” to or that which “may shade over into” rationality of the value and/or instrumental sense, are

> *affectual*: “determined by the actor’s specific affects and feeling states.”
>
> *traditional:* “determined by ingrained habituation.”

Whether or not affectual and traditional action ought to be considered rationales, is a proof I will not offer, but I do not consider their inclusion to be indicative of a model with low explanatory power by virtue of their inclusion alone. Including them here serves the purpose of an axiological analysis or topology of reasoning. As such rational irrationality, may not just be a form emergent from the contradictions of advanced capitalism, the nature of bounded rationality, or the efficacy of heuristics, but also the weighting of reasons as a consistent fact or the weights given in particular.A *topology of rationale,* as a theory/model which is translated into a numerical form may be useful for AI alignment with the “good”.

Correctly trusting AI beyond matters of relatively inert matters will be fruitfully served by a model of rationale, that distinguishes between the various sub-forms of reason(ing): Epistemic, Instrumental, Affectual/Psychoanalytic, Traditional, and Value.

One can set conditions under which a piece of advice of being made for a person, should be outputted as not to be followed , but a regretful machine is not to be reasoned with – in terms of reciprocal discourse. Correctly trusting AIr requires knowing to what extent and why it is to be trusted,so that may mean that one has instrumental and epistemic warrant to say that the profit incentive should never be the primary, and/or exclusive incentive.

Insofar as AI requires a digitised model, AI is a simulation and as a simulation is an instrument. Whilst the model that the simulation runs on may be isomorphic enough to the model before digitisation, there is still a difference.

Trusting AI involves evaluating its fidelity to reality and is dependent on the outcome of a falsification of Alvarado's statement: "Numerical methods the kind used for computer simulations are more often than not guided by the need to reproduce approximate values that only tie them to the original continuous formalities of scientific models but not the phenomena in question (al 8+ ref) In trusting a computer simulation we're trusting a model, not phenomena. Andrew's claim that "We make a category mistake when we claim that a raw mathematical structure lends us predictions or places constraints on what can be observed in nature, and are guilty of reification". "There is no such thing as a solely mathematical account of a target system" (Nguyen and Frigg 2017). "Likewise when we take the existence or qualities of a model to constitute knowledge of the natural world we make a category error and reify the model." (Andrews 2021) This should be a part of a heuristic regarding the use of AI, especially in cases beyond non-conscious empirical and mathematical epistemic applications. Furthermore, if we can make category errors with non-digitised mathematical models, the likelihood and or severity of a mistake increases when mathematical or other models are translated into a numerical scheme which is then translated into a program structure. A relevant change occurs in translation and by default, is significant when that translation is from model to digital program.This is what Alvarado calls discretization. Whilst a series of straight lines can do as a circle it is not a circle. Thus the digital model is not the equation as written and the equation as written is not the phenomenon.It is an abstract representation of an abstract representation.

If we take the FAQs definition of wisdom, it becomes clear that while AI can improve itself through RLHF functions, this doesn't guarantee its ability to recognise that "an old ontology was baking in some problematic assumptions".. In trusting AI we **must** ask ourselves, to what extent is the digital model encompassing the phenomena- whether the numbers alone really are the best depiction of the target system.How to correctly trust AIis a multivectorial question;, each answer is correct in virtue of the scale, domain, frame and aim of the answer space. In this essay I have mostly given descriptions of the constants and constraints specific to specific domains, to give prescriptions that I as a philosopher would be proud of in spite of the asymptotic and context-dependent nature of truth and good, requires addressing particular problems. In so far as a domain general answer can be given wise AI requires what a centaur-like formation of technicians, and philosophers, involved in the interrogation and construction of the simulation, and its outputs with respectto plurality in vectors, domain and scale(instrumental, epistemic, affectual, moral, economic and algorithmic). Consider Theodore Parker's claim that "the arc is a long one, my eye reaches but little ways. I cannot calculate the curve and complete the figure by the experience of sight; I can divine it by conscience. But… I am sure it bends towards justice." We should also remember that any arc has its intentional and unintentional elements, the role of the philosopher seems to be the one reflexive towards models, even the models they use in the reflexive process. Correctly trusting AI requires AI to be produced in spaces where processes and models are rigorously interrogated and integrated along a plurality of vectors.

The question of good models for digital human analogs might benefit from considering paperclip maximisers, Nick Land's machinic desire, Harlon Ellison's AM and the like in light of something like an Aristotlean model of souls as psyches (vegetative, animal and human). Whilst this may seem frivolous it may also be a significant contribution to research and design of material and immaterial artefacts, systems, policy and usage patterns that Don Ihde suggests philosophers can make. Correctly trusting AI requires correctly building it- to serve us, as we are and as we can potentially be. The rationalist tendency in science which reflects philosophy, is useful for the conceptual plumber who must create non-empirical models.This approach must not merely yield ground, but be open to the reciprocal input of so-called irrationalists. This isn't the "anything goes" irrationalism perceived by skim readers of Feyerabend, but rather an understanding of the inherent incompleteness of the rationalist project. Like the positivist project, said incompleteness may be lessened when it is carried out in concert with a sort of rational irrationalism. The differing uses of the word symbolic by

Simon and Bernard Stiegler illustrate this kind of synergy. Simon uses symbolic as that which is easily numerically translated is symbolic. Stiegler sees it exceeding quantified grasp, but both see de-symbolising as an irrational venture for different reasons.Despite their differences, both models help us manage our bounded rationality; Simon posits a limit, for Stiegler symbols are part of an instrumental exteriorisation of memories our own and others. The dispute between the rationalist and the irrationalist, perhaps the analytical and continental, is akin to finding the middle path between the Charybdis of Moloch and the Scylla of Eula for Alexander. For Whitehead, “The divergence between the schools is the quarrel between safety and adventure”, but however we phrase it bi or multi- framework linguality seems to be a virtue for a person to have and indicative of a virtuous institution. Correctly trusting AI requires at minimum the existence of AI that it is correct to trust, multiple dimensions of rationale aside or not. All this has been said with one linguistic Chinese Room argument suppressed: If language is primarily for communication not thought and we wouldnt expect someone devoid of language skills at all to be able to fruitfully contribute to knowledge production compared to someone who has them all things equal, then why would we use a model that suggests seeking wisdom primarily as a quantitative pursuit? If we don’t actually think solely in zeros and ones.

Wei Dai argues that “achieving a good long term future requires getting a lot of philosophical questions right that are hard to answer.”For Dai there are five ways in which AI can go right, and whilst I am not convinced that these problems can be solved through action on hardware, digital models, or software alone, I am sympathetic to the idea that we can mitigate the disaster implied through the utilisation of philosophical expertise in the design of AI of architecture, its models, weights, usage of weighting and policies/laws regarding the use of AI output as epistemic instruments or steerer of a system.

1) What does it mean to solve a philosophical problem? Because of the enormity of the question and without accounting for the varying rationale and their combinations and the implication of a solution's externality, I do not think it can be done well. Especially if philosophical problems have something to do with us as we are**. It may be better to consider a problem approximately and temporarily solved when the solution helps to solve more problems than it creates.** We could also consider solutions to philosophical problems as taking the form of consistent processes or the constants in processes. If Dreyfus is correct, then no answer particularly relevant outside the domain of pure maths, is purely mathematical without viciously bounding ones rationality. The desire to reduce to quantity alone is a pathological irrational rationality.Even if we solve enough of the problems ahead of time, the amount of loss occurred in translation to numerical form in most cases, will prohibit us from correctly trusting AI for all purposes in all domains over multiple integrated sources.

2) Insofar as bounded rationality captures something essential about the human condition, understanding philosophical reasoning as well as we understand mathematical reasoning is no guarantee that the translation of this understanding will capture all of it.

3) If to be human is to be boundedly rational and therefore somewhat irrational/suboptimal then programming AI to learn philosophical reasoning from humans or using human simulations is asking for, for code to be irrational. Yet if we are to try to induce the more irrational elements of human behaviour into AI or for it to do a better job of capturing them in simulation we must have better models of behaviour, which requires philosophers of emotion, phenomenology, psychoanalysis and social ontology to be co-architects (with scientists and computer scientists) of the models to be digitised.

What is philosophy and can AI solve it?

For Kenneth L. Pearce philosophy is “the use of logic in the critical examination of one’s most deeply held beliefs & assumptions”

For Heidegger it’s “useless, though sovereign, knowledge.”

For Wittgenstein it’s “a battle against the bewitchment of our intelligence by means of language.”

For Whitehead its use is to “maintain an active novelty of fundamental ideas illuminating the social system.” He also distinguishes it from poetry saying that whilst both “seek to express that ultimate good sense which we term civilization..poetry allies itself more to metre, philosophy to mathematic pattern.”

For Alvarado and Symons “on our view it is crucial to recognize that scientific inquiry is not a basic epistemic practice but rather a very special cultural practice that is designed, in part, to overcome the evident limitations of our ordinary epistemic conditions” and that “Philosophy is another social practice that sets abnormally high epistemic standards. In our case, we aim high with respect to what should count as a rationally persuasive argument”

Whilst Wei Dai is correct to say that “given philosophical difficulties, the target we’d have to hit with AI is even smaller than it might otherwise appear.” In truth, there are no philosophical problems, just concepts made and used in better and worse ways than others.

Philosophy is difficult, aside from the masturbatory impulse, that may be why Heidegger also said that “philosophizing involves the possibility of a challenge.” And that is why AI can help, but not finish philosophy.

Conclusion

Whilst I am sheepish towards the work of philosophers as more than historians of ideas, and compared but not synthesised for a use other than filling books, I do hope that in a world of increasing digital computation, the role of the philosopher will not be like that of *in vivo* birth described in Shulamith Firestone's *Dialectic of Sex*: “eventually acknowledged as clumsy, inefficient, and painful, would be indulged in, if not at all, only as a tongue in cheeky archaism, just as already women today wear virginal white to their weddings.” (Firestone 1970). I think we’d be worse for it. I also hope that we will not function as Hegel’s Owl of Minerva that flies after dusk, but rather more of what Don Ihde called the “Hemingway role”, in reference to this role during the Spanish civil war as a member of the ambulance corps, Hemingway had to practise “triage on the spot.” Moving forward, instead of the role of historian or after-the-fact polemicist or even that of Hemingway in all contexts, the philosopher could move into what Ihde called the “R&D role” where they can focus not on norms, but exploring problems and possibilities in an epistemological context.

Expertise in policy and machine learning alone is not enough to guarantee widespread AI does not cement us in the basis of bad governance, just as philosophical expertise is not a guarantee of a position in the basin of good governance. Both are necessary constants and amplifiers in a system hoping for positive attraction. Correctly trusting AI, in part, requires

acknowledging that an *in silico* model which begets *in silico* modelling, process and output - which may beget an *in vivo* output - are derived from an *in stilo* or *in cogito* model. All are varieties of the way in which something can be *in situ*. The role of the philosopher may be to triage, plurally, *in situ*. An AI that is correct to trust and correctly trusted is more likely to be so when philosophers are promoted from the Hemmingway role to an R&D role. Alignment, like all concepts, is multi-stable, and Dreyfus was right to say whatever we are, a purely numerical model is more incomplete than one that involves non-empirical elements and constants.

Creating wise AI requires respecting the fact that the idea that things can be reduced solely to numerical forms is relatively recent, and that such a reduction is not proof of its multistable-multidomain-multiaim correctness. Until coding of digital computers evolves beyond numerical level, the best that can be done, for the sake of mitigating the possible disasters of enclosure within the basin of bad unreflexive governance, involves collaboration and reciprocity between those who make, test, and implement models: *in stilo, silico* and *vivo*.

Bibliography

Alvarado, Ramón. 2021. “Computer Simulations as Scientific Instruments.” *Foundations of Science* 27 (3): 1183–1205. https://doi.org/10.1007/s10699-021-09812-2.

———. 2022. “What Kind of Trust Does AI Deserve, if Any?” *AI And Ethics* 3 (4): 1169–83. https://doi.org/10.1007/s43681-022-00224-x.

Andrews, Mel. 2021. “The Math Is Not the Territory: Navigating the Free Energy Principle.” *Biology & Philosophy* 36 (3). https://doi.org/10.1007/s10539-021-09807-0.

Beal, Bree. 2022. “The Polyphony Principle.” *Behavioral and Brain Sciences* 45 (January). https://doi.org/10.1017/s0140525x2200108x.

Beer, Stafford. 2002. “What Is Cybernetics?” *Kybernetes* 31 (2): 209–19. https://doi.org/10.1108/03684920210417283.

Bermúdez, José Luis. 2022a. “Rational Framing Effects: A Multidisciplinary Case.” *Behavioral and Brain Sciences* 45 (January). https://doi.org/10.1017/s0140525x2200005x.

———. 2022b. “Rational Framing Effects: A Multidisciplinary Case.” *Behavioral and Brain Sciences* 45 (January). https://doi.org/10.1017/s0140525x2200005x.

Dreyfus, Hubert L., and Stuart E. Dreyfus. 1991. “Making a Mind Versus Modelling the Brain: Artificial Intelligence Back at the Branchpoint.” In *Springer eBooks*, 33–54. https://doi.org/10.1007/978-1-4471-1776-6_3.

Durán, Juan M. 2017. “Varieties of Simulations: From the Analogue to the Digital.” In *Springer eBooks*, 175–92. https://doi.org/10.1007/978-3-319-55762-5_12.

Franklin, Ursula. 1999. *The Real World of Technology*. House of Anansi.

Friedman, Milton, and Rose D. Friedman. 2002. *Capitalism and Freedom: Fortieth Anniversary Edition*. University of Chicago Press.

Gigerenzer, Gerd. 2010. “Moral Satisficing: Rethinking Moral Behavior as Bounded Rationality.” *Topics in Cognitive Science* 2 (3): 528–54. https://doi.org/10.1111/j.1756-8765.2010.01094.x.

Hartmann, Stephan. 1996. “The World as a Process: Simulations in the Natural and Social Sciences.” *Institute of Philosophy*, January. http://philsci-archive.pitt.edu/2412/1/Simulations.pdf.

Havercroft, Jonathan, and David Owen. 2016. “Soul-Blindness, Police Orders and Black Lives Matter.” *Political Theory* 44 (6): 739–63. https://doi.org/10.1177/0090591716657857.

Hegselmann, R., Ulrich Mueller, and Klaus G. Troitzsch. 2013. *Modelling and Simulation in the Social Sciences From the Philosophy of Science Point of View*. Springer Science & Business Media.

Heidegger, Martin. 2005. *The Essence of Human Freedom: An Introduction to Philosophy*. A&C Black.

———. 2012. *Contributions to Philosophy (of the Event)*. Indiana University Press.

Hirji, Sukaina. 2018. “What’s Aristotelian About neo‐Aristotelian Virtue Ethics?” *Philosophy and Phenomenological Research* 98 (3): 671–96. https://doi.org/10.1111/phpr.12520.

Hubig, Christoph, and Andreas Kaminski. 2017. “Outlines of a Pragmatic Theory of Truth and Error in Computer Simulation.” In *Springer eBooks*, 121–36. https://doi.org/10.1007/978-3-319-55762-5_9.

Huemer, Michael. n.d. “Irrational Politics.” https://spot.colorado.edu/~huemer/papers/irrationality.htm.

Humphreys, Paul. 2004. *Extending Ourselves: Computational Science, Empiricism, and*

Scientific Method.

Ihde, Don. 1999. “Technology and Prognostic Predicaments.” *AI & Society* 13 (1–2): 44–51. https://doi.org/10.1007/bf01205256.

Ladyman, James, Don Ross, David Spurrett, and John Collier. 2007a. *Every Thing Must Go: Metaphysics Naturalized*. OUP Oxford.

———. 2007b. *Every Thing Must Go: Metaphysics Naturalized*. OUP Oxford.

Lenhard, Johannes. 2017. “The Demon’s Fallacy: Simulation Modeling and a New Style of Reasoning.” In *Springer eBooks*, 137–51. https://doi.org/10.1007/978-3-319-55762-5_10.

Marcuse, Herbert. 1991. *One-Dimensional Man: Studies in the Ideology of Advanced Industrial Society*. 2nd ed. Routledge.

Midgley, Mary. 2000. *Utopias, Dolphins and Computers: Problems of Philosophical Plumbing*. Psychology Press.

Nguyen, James, and Roman Frigg. 2017. “Mathematics Is Not the Only Language in the Book of Nature.” *Synthese* 198 (S24): 5941–62. https://doi.org/10.1007/s11229-017-1526-5.

Resch, Michael M. 2017. “On The Missing Coherent Theory of Simulation.” In *Springer eBooks*, 23–32. https://doi.org/10.1007/978-3-319-55762-5_3.

Resch, Michael M., Andreas Kaminski, and Petra Gehring. 2017. *The Science and Art of Simulation I: Exploring - Understanding - Knowing*. Springer.

Saam, Nicole J. 2017. “Understanding Social Science Simulations: Distinguishing Two Categories of Simulations.” In *Springer eBooks*, 67–84. https://doi.org/10.1007/978-3-319-55762-5_6.

Schumacher. 1982. “E. F. Schumacher: on Technology for a Democratic Society.” In *Small Is Possible*, by George McRobie, 1–18. Sphere Books.

Simon, Herbert A. 2019. *The Sciences of the Artificial, Reissue of the Third Edition With a New Introduction by John Laird*. MIT Press.

Simon, Herbert Alexander. 1957. *Models of Man: Social and Rational; Mathematical Essays*

on Rational Human Behavior in Society Setting. New York : Wiley.

Sinding-Larsen, Henrik. 1991. “Computers, Musical Notation and the Externalisation of Knowledge: Towards a Comparative Study in the History of Information Technology.” In *Springer eBooks*, 101–25. https://doi.org/10.1007/978-1-4471-1776-6_7.

Stiegler, Bernard. 2013. *What Makes Life Worth Living: On Pharmacology*. Polity.

Symons, John, and Ramón Alvarado. 2019. “Epistemic Entitlements and the Practice of Computer Simulation.” *Minds and Machines* 29 (1): 37–60. https://doi.org/10.1007/s11023-018-9487-0.

Weber, Max. 1978. *Economy and Society: An Outline of Interpretive Sociology*. Univ of California Press.

Wellner, Galit. 2020. “The Multiplicity of Multistabilities: Turning Multistability Into a Multistable Concept.” In *Philosophy of Engineering and Technology*, 105–22. https://doi.org/10.1007/978-3-030-35967-6_7.

Whitehead, Alfred North. 1968. *Modes of Thought*. http://www.gpmcf.org/PDFs/cornerdec2015.pdf.

Yudkowsky, Eliezer. 2018. *Map And Territory*.