minimal space is voided by the very criteria that are proposed to unbind posthuman intelligence from human agency. The room reserved for a level of interpretability in order to prevent the collapse of unbound posthumans into unintelligible alienness shrinks to nothing. And just as the line between the 'strangeness' of unbounded posthumans and unintelligible 'radical alienness' blurs, so does the distinction between the disconnection thesis and speculative apophatic theology.

Finally and most importantly, what is missing in this critique is a defence of rational agency against both parochial humanism and unbounded posthumanism. But to properly engage with the artificial potencies of rational agency, we have to look at the structure and functions of language, its generative computational architecture, which stretches over its syntactic, semantic, and pragmatic aspects. This is something that we will closely investigate in chapters 6 and 7, which deal with the second stage in our thought experiment. However, it should be noted that a definition of general intelligence in terms of semantic complexity and a rationalist critique of the myths of superintelligence are not by themselves sufficient. One needs to demonstrate that the underlying premises and methods presupposed by such scenarios are not merely inadequate but logically flawed on the basis of their own assumptions. The equation of general intelligence with compression, the valorisation of purely inductivist methods in cognitive science,[93] and the program of artificial general intelligence as universal methods capable of bypassing the problems of epistemic multimodality and semantic complexity are among the presuppositions which lead to the myths of omniscient or omnipotent intelligence.

Those narratives of superintelligence that make up the majority of views and hypotheses about AGI are deeply enmeshed in notions whose supposedly inherent association with strong forms of AI is far from self-evident: personal autonomy, value appraisal and revision, organised goal-seeking and self-enhancement. Each of these presuppose forms of self-knowledge that enable and incite purposeful action and deliberate interaction: negotiation,

[93] For a critique of purely inductivist trends in the philosophy of mind and cognitive sciences, see the Appendix.

persuasion, or even threat and plotting. So are these narratives of AGI simply exercises in anthropomorphism? Or is it rather the case that a strong AI's complex capacities for action and cognition would be subject to certain necessary conditions of realization, that they would require some more basic capacities and faculties functionally isomorphic to those competencies that undergird *our* complex cognitive-practical abilities? If this is the case, then what are these necessary conditions, and do they have other ramifications and entail other capacities and qualities than those we have included in our AGI scenarios to date? If so, then the various accounts of malevolent, benevolent, and disconnected AI should not be taken as anything more than speculative fabulations; the extreme nature of these scenarios is precisely an artefact of our ignorance with respect to those necessary constraints and conditions, the ramifications of which would both complicate and govern the behaviour of a strong AI.

HARD VS. SOFT PAROCHIALISM

In opposition to the singularity myths of superintelligence, the AS-AI-TP thought experiment attempts to accomplish three tasks. Firstly, to highlight the conditions necessary for the realization of basic capacities which are prerequisites for the development of those complex forms of action and cognition commonly attributed to human-level AI. Secondly, to examine the ramification and generative entrenchment of developmental constraints attached to the conditions that permit higher-order abilities to emerge. Thirdly, to explore the consequences arising from the exercise of these higher-order—theoretical and practical—cognitions. If an AGI has at the very least all of our cognitive capacities, it is as strongly attached as we ourselves are to the conditions necessary for the realization of complex cognitive abilities. And if the initial capacities of AGI share this common ground with our own intelligence, then this will affect our assessment of how far a self-augmenting AGI can diverge from us toward extremes of malevolence, benevolence, or disconnection from humans. In other words, these necessary conditions should be thought of as constraints that simultaneously make the realization of higher-order abilities possible and limit the ways in which they behave or can be artificially realized—much like the concept of boundary conditions for the analysis of a system's tendencies.

Accordingly, the thought experiment we will set out below is an argument about the conceptual problems involved in the construction of an AGI. In a roundabout manner, the thought experiment also addresses a more fundamental question about modelling AGI on humans and what it would entail for us to be the models of something that should have, at least, all of our abilities. Alternatively, this question can be formulated as follows: What kinds of revisions or corrections must our self-conception go through in order for us to be able to formulate a nontrivial conception of what AGI is and what it can be? Or, in simplified form, it can be framed as follows: *Should AGI converge upon humans or should it diverge from them?*

The answer to this question depends upon a number of presuppositions: the level of generality in General Intelligence, what we mean by the human, and whether the question of mirroring or artificial realization and divergence is posed at the level of functional capacities, that of structural constitution, that of the methodological requirements necessary for the construction of AGI, or that of the diachronic consequences of its realization.

If we are parochially limiting the concept of the human to certain local and contingently posited conditions—namely, a specific structure or biological substrate and a particular local transcendental structure of experience—then the answer must be divergence. Those who limit the significance of the human to this parochial picture are exactly those who advance parochial conceptions of AGI. There is a story here about how anti-AGI sceptics and proponents of parochial conceptions of AGI are actually two sides of the same coin. On the one hand, there are those who think biological structure or the structure of human experience are foreclosed to artificial realizability. On the other, there are those who think models constructed on a prevalent 'sentient' conception of intelligence, inductive information processing, Bayesian inference, problem-solving, or emulation of the physical substrate are *sufficient* for the realization of AGI. The positions of both camps originate in a deeply conservative picture of the human which is entrenched either in biological chauvinism or in a provincial account of subjectivity, a mystical privileging of the human's lived experience or a dogmatic adherence to the abstractly universal laws of thought as, ultimately, the laws of nature.

The only thing that separates them is their strategy with regard to their base ideological assumption: the sceptics inflate this picture into a rigid anthropocentricism, whereas the proponents of parochial AGI attempt to maximally deflate it. Thus we arrive at either a thick notion of general intelligence that does not admit of artificial realizability, or a notion of general intelligence too diluted for it to have any classificatory, descriptive, or theoretical import with regard to what intelligence is or, more specifically, what human-level intelligence would entail. In the latter case, the concept of general intelligence is watered down to prevalent yet rudimentary intelligent behaviours based on the assumption that the difference between general intelligence and mere intelligent behaviours prevalent in nature is simply quantitative.

Conceptualizing activities or, more broadly, theoretical and practical cognitions, are taken to be pattern-governed activities, and to the extent that nature is replete with unexceptional pattern-governed behaviours, conceptual cognition or human activities are then treated at the same level as any other such behaviour. But, as Wilfrid Sellars points out, although the conceptual activities that underline the exceptionality of the human may indeed be pattern-governed behaviours, they are not just any sort of patterns. They are pattern-governed behaviours that are *sui generis* because they are properly speaking rule-governed—that is to say, because they have a formal autonomy that arises from their functioning according to intra-pattern-governed *norms* of behaviour (i.e., *rules* of transition or inference). But conceptual activities are also *sui generis* in a stronger sense: their formal autonomy, which is logical and linguistic, enables the recognition of any other pattern-governed behaviour in nature. In other words, without the exceptionality of pattern-governed conceptual activities qua rules, the issue of the nonexceptional nature of the human within the universe or the equivocation around pattern-governed activities wouldn't even arise in the first place.

It is one thing to explain the causal origins of thinking, as science commendably does; it is an entirely different thing to conflate thinking in its formal or rule-governed dimension with its evolutionary genesis. Being conditioned is not the same as being constituted. Such a conflation not only sophistically elides the distinction between the substantive and the formal, it also falls victim to a dogmatic metaphysics that is impulsively blind to its own epistemological and methodological bases qua origins.

It is this genetic fallacy that sanctions the demotion of general intelligence as qualitatively distinct to a mere quantitative account of intelligent behaviours prevalent in nature. It should not come as any surprise that this is exactly the jaded gesture of antihumanism upon whose shoddy pillars today's discourse of posthumanism supports its case. Talk of thinking forests, rocks, worn shoes, and ethereal beings goes hand in hand with the cult of technological singularity, musings on Skynet or the Market as speculative posthuman intelligence, and computers endowed with intellectual intuition. And again, by now it should have become obvious that, despite the seeming antagonism between these two camps—one promoting the so-called egalitarianism of going beyond human conditions by dispensing with the rational resources of critique, the other advancing the speculative aspects of posthuman supremacy on the grounds of the technological overcoming of the human condition—they both in fact belong to the arsenal of today's neoliberal capitalism in its full-on assault on any account of intelligence that may remotely insinuate an ambition for collective rationality and imagination.

Having dispensed with the categorical distinctions between various pattern-governed behaviours and conceptual activities as yet another set of trivialized and unexceptional natural processes, the proponent of parochial AI then concludes: If we artificially realize and put together enough rudimentary behaviours and abilities, we will essentially obtain general intelligence. In other words, the trick in realizing general intelligence is to abstract basic abilities from below and then find a way to integrate and artificially realize them. Let us call this approach to the AGI problem *hard parochialism*. Hard parochialists tend to overemphasize the prevalence of intelligent behaviours and their sufficiency for general intelligence, and become heavily invested in various panpsychist, pancomputationalist, and uncritically anti-anthropocentric ideologies that serve to justify their theoretical commitments and methodologies.

On the other hand, if we define the human in terms of cognitive and practical abilities that are minimal yet *necessary* conditions for the possibility of any scenario that involves a sustained and organized self-transformation (i.e., value appraisal, purposeful decision, and action based on knowledge that harbours the possibility of deepening its own descriptive-explanatory powers), and deliberate interaction (i.e., negotiation, persuasion, or even threat and plotting), then the answer must be functional mirroring, despite structural divergence.

But then a different question arises: Should we limit the model of AGI—the hermeneutics of general intelligence—to the functional mirroring of the capacities and abilities of human agency?

My answer to this question is an emphatic No. Functional mirroring or convergence is a *soft parochialist* approach to the problem of AGI and the question of general intelligence. In contrast to hard parochialism, functional mirroring or convergence upon the human is necessary for grappling with the conceptual question of general intelligence as well as the modelling and methodological requirements for the construction of AGI. But even though it is necessary, it is not sufficient. It has to be coupled with a critical project that can provide us with a model of experience that is not restricted to a predetermined transcendental structure and its local and contingent characteristics. In other words, it needs to be conjoined with a critique of the transcendental structure of the constituted subject (existing humans).

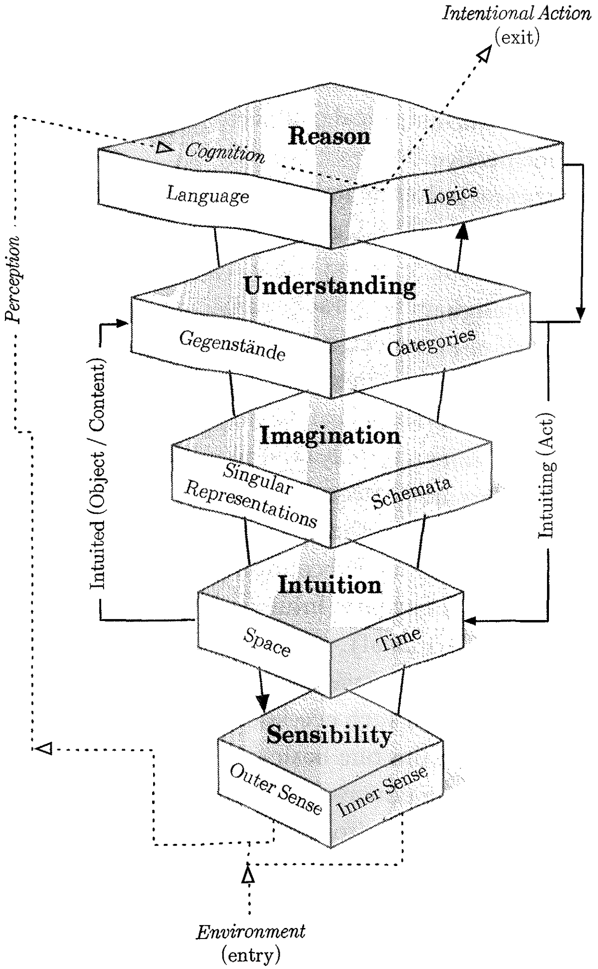
In limiting the model of AGI to the replication of the conditions and capacities necessary for the realization of human cognitive and practical abilities, we risk reproducing or preserving those features and characteristics of human experience that are purely local and contingent. We therefore risk falling back on the very parochial picture of the human as a model of AGI that we set out to escape. So long as we leave the transcendental structure of our experience unquestioned and intact, so long as we treat it as an essence, we will gain inadequate objective traction on the question of what the human is and how to model an AGI that is not circumscribed by the contingent characteristics of human experience. But why is the critique of the transcendental structure indispensable? Because the limits of our empirical and phenomenological perspectives with regard to the phenomena we seek to study are set by transcendental structures. Put differently, the limits of the objective description of the human in the world are determined by the transcendental structure of our own experience. The limits of the scientific-empirical perspective are set by the limits of the transcendental perspective.[94]

But what are these transcendental structures? They include any and all of the structures—physiological (e.g., the locomotor system and neurological mechanisms), linguistic (e.g., expressive resources and internal logical structure of natural languages), paradigmatic (e.g., frameworks of theory-building in sciences), or historical, economic, cultural, and political—that regulate and canalize our experience. These transcendental structures need not be seen separately, but instead can be mapped as a nested hierarchy of interconnected and at times mutually reinforcing structures that simultaneously constitute, regulate, and constrain experience. If we were to imagine a Kantian-Hegelian diagram of this nested hierarchical structure, it would be represented by a nested hierarchy of conditions and faculties necessary for the possibility of mind: [Sensibility [Intuition [Imagination [Understanding [Reason]]]]] (see diagram overleaf).

Transcendental structures then would be outlined as structures required not only for the realization of such necessary conditions and faculties, but also for moving upward from one basic condition to a more composite condition as well as moving downward from complex faculties to harness the power of more basic faculties (for example, deployments of the concept in order to manipulate the imagination in its Kantian sense—the function of the productive imagination, which is simply understanding in a new guise).

In so far as any experience is perspectival, and this perspectival character is ultimately rooted in transcendental structures, any account of intelligence or general intelligence is circumscribed by the implicit constraints of the transcendental structure of our own experience. Regardless of whether or not we model AGI on humans, our conceptual and empirical descriptions

[94] I owe this insight to Gabriel Catren, whose work has been pivotal for me in building this critique and arriving at conclusions which, however, may stray from the sound conclusions reached by his meticulous analyses.



Kant's dimensionally-varied conditions of possibility for having mind: In this functional schema, not only is each level in interaction with other levels (e.g., categories of understanding are regulative of intuition and constitutive of experience), the hierarchy itself is in interaction with the environment, whether in the guise of affects upon sensations (from the bottom) or communication with other 'persons' (from the top). This interactionist multilevel view of conditions of possibility captures three groups of fundamental relations intrinsic to mind, between (1) metaphysics (universal categories as predicates of objects qua appearances), (2) logic (forms of judgments), and (3) psychology (sensible-perceptual synthesis).

of what we take to be a candidate model of general intelligence are always implicitly constrained by our own particular transcendental structures. Here I am not endorsing the view that we should model a hypothetical AGI on something extra-cognitive or something other than the human mind. Whatever model of AGI we come up with will inevitably be modelled on the human mind or, more specifically, on the a priori acts of cognition (*erkenntnis*) and the oughts of our theoretical and practical reason. This inexorable recourse to the a priori dimensions of the human mind is not what I am criticizing, for it is the only necessary and sound way to handle the problem of AGI. Anything else will be a hopeless shambles of dogmatic metaphysics, a whimsical cabinet of curiosities luring the benighted cult of posthumanism to speculate endlessly about its magical qualities.

Rather, the critique takes aim at the idea that the categories of the conceptualizing mind, the pure concepts of understanding, are bound up with the local and contingent structure of experience. To the extent that we employ these categories to give structure to the world (the universe of data) and to make sense of the experience of who we are in the world, and furthermore, in so far as the extent to which the a priori categories are entangled with the contingent aspects of experience is still a widely unexamined issue, the critique of our particular transcendental structures should be treated as nothing more or less than the extension of critical philosophy. Even though it is now science that can carry the banner of this critique in the most rigorous way, it remains a genuine continuation of the gesture initiated by critical philosophy. Furthermore, the critique of the transcendental structure is in reality nothing but the fomentation of the Hegelian gesture of disenthralling reason from the residual influence of Kantian conservatism for which experience and reason are still muddled together.

Modelling AGI on the transcendental structure of our experience in the sense outlined above is in fact a form of anthropocentrism that is all the more insidious to the extent that it is hidden, because we take it for granted as something essential and natural in the constitution of human intelligence and our experience of it. In leaving these transcendental structures intact and unchallenged, we are inevitability liable to reinscribe them in our objective model. Anti-anthropocentric models of general intelligence and those philosophies of posthuman intelligence that have anti-humanist commitments are particularly susceptible to the traps of this hidden form of essentialism. Because by treating the rational category of sapience as irrelevant or obsolete, and by dispensing with the problem of the transcendental structure as a paltry human concern, we become oblivious to the extent to which our objective conceptual and empirical perspectives are predetermined by our transcendental structure. In remaining oblivious to the problem of transcendental blind spots, we place ourselves at far greater risk of smuggling in essentialist anthropocentrism, replicating the local and contingent characteristics of human experience in what we think is a radical nonanthropocentric model of general intelligence. It is those who discard what nontrivially distinguishes the human that end up preserving the trivial characteristics of the human in a narrow conception of general intelligence.

The above argument can be reformulated in the context of the necessary correspondence between intelligence and the intelligible as provided in chapter 1. Intelligence is an illusion if it is disconnected from the labour of intelligibility and thus from the requirements or positive constraints which enable it to engage with the intelligible, including its own intelligibility. Dispensing with such constraints can only effectuate a conception of intelligence that is a reservoir of human subjective biases and personal flights of fancy. But at the same time, if we are serious about a broader conception of intelligence that differs from our impression of intelligence here and now, we should think about how such local and evolutionarily given constraints can be modified so that the concept of intelligence can be reimagined or reinvented according to a more expansive idea of an intelligible universe.

It is of course not the case that AGI research programs must wait for a thoroughgoing critique of the transcendental structure to be carried out via physics, cognitive science, theoretical computer science, or politics before they attempt to put forward an adequate model; the two ought to be understood as parallel and overlapping projects. In this schema, the program of the artificial realization of the human's cognitive-practical abilities coincides with the project of the fundamental alienation of the human subject, which is precisely the continuation and elaboration of the Copernican enlightenment, moving from a particular perspective or local frame to a perspective or experience that is no longer uniquely determined by a particular and contingently constituted transcendental structure. In the same vein, the project of artificial general intelligence, rather than championing singularity or some equally dubious conception of the technological saviour, becomes a natural extension of the human's process of self-discovery through which the last vestiges of essentialism are washed away. What remains after this process of retrospective reassessment and prospective revision may indeed—as Roden suggests—bear no resemblance to the manifest self-portrait of the human in which our experience of what it means to be human is anchored.

However, the precipitate abandonment of this manifest self-portrait is a sure way to reentrench the very prejudices embedded within it. We may indeed arrive at a conception of posthuman intelligence that is in no sense in congruity with what we take ourselves as, here and now. But it is highly contentious and unwarranted to claim that we can arrive at such a conception of intelligence absent or despite what we take ourselves to be here and now. As indicated above, such a speculation about future intelligence inevitably degenerates into negative theology. Genuine speculation about posthuman intelligence begins with the suspension (*aufhebung*) of what we *immediately* appear to ourselves to be. It is thus the product of an extensive labour of determinate negation that cannot start from nowhere and nowhen, but can only begin with the determination of a conception of ourselves at the historical juncture within which we recognize and make judgements about ourselves, i.e., a definite where and when. To arrive at a view of intelligence from nowhere and nowhen we can therefore only begin with a critical and objective view on the where and when of what we take the human to be. That is to say, a nontrivial conception of artificial general intelligence rests on our own adequate self-conception as a task—one that is revisable, self-critical, and by no means taken for granted as immediate or a completed totality.

The structural-functional analysis of the conditions and capacities necessary for the realization of human cognitive-practical abilities is thus an obligatory framework for AGI research. But the sufficiency of this framework depends upon how far we deepen our investigation into the transcendental structure of human experience and how successful we are in liberating the model of the human subject (or agent) from the contingent characteristics of its experience. In this sense, a consequential paradigm of AGI should be seen as the convergence of two projects:

> (1) Examination of the conditions and capacities necessary for the realization of what, for now, we can call the human mind, as well as the more applied question of how to artificially realize them.
>
> (2) Critical investigation into the transcendental structure of experience in order to develop a different model of experience that is no longer treated as essential or foundationally given—that is, one that is no longer fixated upon a particular local and contingently framed transcendental structure.

Thus, to the question of whether AGI should be modelled on humans or not, and if so on what level, we answer as follows: AGI should be modelled on the human in the sense that it should functionally converge on the conditions and capacities necessary for the realization of human cognitive-practical abilities. But it should diverge from the transcendental structure of the constituted human subject. However, the success of this divergence depends upon (1) our success in rationally-scientifically challenging the given facts of our own experience and in doing so reinventing the figure of the human—ourselves—beyond strictly local transcendental structures and their contingent characteristics (this is the project of the fundamental alienation of the human), and (2) the success of AGI research programs in extending their scope beyond applied dimensions and narrow implementation problems towards theoretical problems that have long vexed physics, cognitive science, and philosophy.

Modelling AGI on human agency is not merely a strategy for tackling the conceptual problems involved in constructing a nonparochial artificial intelligence, but also more fundamentally a strategy for coming to grips with questions concerning the nature of minds, what they are, what they can become, and what they can do. If we posit ourselves as a model of an artificial agency that has all the abilities that we have, then we ought to examine exactly what it means for us to be the model for that which harbours the possibility of being—in the broadest sense—better than us. This is the question of modelling future intelligence on something whose very limits can be perpetually renegotiated—that is, a conception of human agency not as a fixed or settled creature but as a theoretical and practical life form distinguished by its ability to conceive and transform itself differently, by its striving for self-transformation in accordance with the revisable conception it has of itself. Ultimately, the question of what the mind is and what it can do is a matter of developing a project in which our process of self-discovery and self-transformation fully overlaps and in a sense reinforces the programme for the realization of an agency that can outstrip what we manifestly conceive ourselves as, here and now. It is within the ambit of this project—the project of inquiring into the meaning and possibilities of agency, by at once identifying and severing all *essentialist* attachments of this meaning or possibility to a particular local or contingent structure—that the human and AGI become non-tautological synonyms. The nontrivial meaning of the human lies in its ability to revise and transform itself, its ability to explore what the human is and what it can become. The nonparochial conception of AGI is simply the continuation and realization of this meaning in its pure or autonomous form.

The opposition between the possibility of a thinking machine and the actuality of the human agent should be exposed as a false dichotomy that can only be precariously maintained within the bounds of an essentialist interpretation of the mind as necessarily attached to a particular local or contingent transcendental structure. To put it more tersely, the source of this false dichotomy lies precisely in mistaking the local and contingent aspects of experience for universal and necessary acts of cognition, the particular conditions of the former for the general conditions of the latter. To reject and break away from this false dichotomy in all its manifestations, it is necessary to fully distinguish and unbind reason (the labour of conception) from subjectivist experience. This is not to dispense with the significance of experience in favour of a contentless abstract account of reason. It is rather the condition necessary for reassessing the extant categories—the general concepts by virtue of which we can have experience in the first place, the structures which render the world and our experience of it intelligible. The unbinding of reason from experience is a required step in order to expand and reshape our experience beyond what is manifestly essential or supposedly given to us. This Hegelian program neither impoverishes reason nor disposes of the significance of experience, but instead opens the way for the fully charged vector of cognitive progress that Nicholas Rescher attributes to science. Here the logical sophistication of the conceptual-inferential resources of the theory and the enlargement of the field of possible experience lie side-by-side in an imbalanced configuration whose very fragility guarantees the dynamic complexity of scientific inquiry:

> For rational beings will of course try simple things first and thereafter be driven step by step toward an ever-enhanced complexification. In the course of rational inquiry we try the simple solutions first, and only thereafter, if and when they cease to work—when they are ruled out by further findings (by some further influx of coordinating information)—do we move on to the more complex. Things go along smoothly until an oversimple solution becomes *destabilized by enlarged experience*. For a time we get by with the comparatively simpler options—until the expanding information about the world's *modus operandi* made possible by enhanced new means of observation and experimentation insists otherwise. And with the expansion of knowledge those new accessions make ever increasing demands. And so evolution, be it natural or rational—whether of animal species or of literary genres—ongoingly confronts us with products of greater and greater complexity. Man's cognitive efforts in the development of natural science manifests a Manichaean-style struggle between complexity and simplicity—between the impetus to comprehensiveness (amplitude) and the impetus to system (economy).[95]

The critique of transcendental structures is strictly a collective project comprised of procedural methods and incremental tasks. On a groundwork

[95] N. Rescher, *Epistemology: An Introduction to the Theory of Knowledge* (Albany, NY: SUNY Press, 2003), 235 (emphasis mine).

level, it begins theoretically by distinguishing necessary conditions for the constitution of theoretical and practical agency from contingent aspects of the subject's constitution, characteristics of an objective reality from characteristics of the subject's experience. At this stage, the critique tackles two fundamental overlapping questions: the extent to which objective descriptions of reality at various levels (our world-structuring categories) are biased or distorted by the contingent characteristics of our experience, and the extent to which the exercise of our theoretical and practical abilities is caught up in or determined by the contingent positioning of our particular transcendental structures (be they associated with our terrestrial habitat, neurophysical systems, cultural environment, family, gender, economy, etc.).

On the basis of this theoretical phase, the project then proceeds to inquire into the possibilities of transforming and diversifying the transcendental structures of agency, renegotiating the extant categories through which we understand ourselves and our position in the world. This is an experimental phase in which the possibilities of transcendental variation, and thus the possibilities of releasing experience—and by extension the theoretical and practical abilities thereby made possible—from limitative attachments to any unique or allegedly essential local transcendental structure are examined. The central task of this stage is to expand the range and type of abilities by altering and reorganizing transcendental structures. Once the prospects of varying transcendental structures and transformation of abilities are systematically outlined and evaluated, the project shifts toward the applied dimension, that of developing implementable mechanisms and systems that can support the realization of new abilities by either modifying or replacing the transcendental structures of the constituted subject.

What begins as a systematic theoretical inquiry into the limits and regulative regimes of the transcendental structures of the constituted subject evolves into an applied system for the transformation of the subject and the maximization of agency. Thus understood, the critique of transcendental structures is the compass of self-conception and self-transformation. By challenging the established characteristics of our experience of ourselves in the world and by renegotiating the limits postulated by our contingent confounding the contingent characteristics of human experience with the conditions necessary for the realization of human abilities (acts of *erkenntnis*), and thus threatens to relapse into a hard parochialist approach to the questions of what general intelligence is and how it can be artificially realized. If we treat the human as a model of AGI, then this model should not only be a model by which we can identify and differentiate the conditions and capacities necessary for the realization of theoretical and practical cognitions, but also a model within which we can renegotiate the characteristics of general intelligence by renegotiating the limits and characteristics of human experience. This is the *outside view of ourselves as a toy model AGI*: A position that allows us to treat ourselves—both our functional capacities and what we take ourselves *as*—from an objective point of view, a view from nowhere and nowhen. This is a viewpoint that distinguishes the necessary conditions and capacities for the realization of the theoretical and practical faculties, or engagement with the intelligible, but which at the same time is not bound in principle to the local characteristics of the subject's experience. Within the scope of this viewpoint, the human is a toy model or construction kit for AGI as much as artificial general intelligence is a constructive model for exploring the human.

But what exactly is a 'toy model'? Toy models are simplified or compressed models that are capable of accommodating a wide range of theoretical assumptions for the purpose of organizing and constructing overarching narratives (or explicit metatheories) that change the standard and implicit metatheoretical interpretations according to which such theoretical items are generally represented. In other words, by explicitly changing the metatheoretical narrative, toy models provide new interpretations of problems and puzzles associated with the implicit metatheoretical frameworks within which theoretical ideas and observations are interpreted. To this end, a rigorous and internally consistent toy model can offer insights about how to solve these puzzles or how to overcome the setbacks caused by the standard interpretations. What separates toy models from models is not just that they are simplified enough to enable us to tinker with the internal theoretical structure of a model, but that they are explicit metatheories. All theories are metatheories, but within regular theoretical models metatheoretical assumptions are usually implicit or hidden, whereas toy models are *explicitly* metatheoretical and in fact the simplification (what gives them the name 'toy', a tinkermodel) serves as a strategy for bringing hidden metatheoretical assumptions out into the open by tinkering with the internal variables of the model without getting bogged down in theoretical details.

Therefore in reality a toy model is a model capable of making explicit its implicit metatheoretical assumptions. And what are these implicit metatheoretical assumptions? They are precisely the implicit or hidden assumptions that arise from applying the characteristics of our subjective experience to our objective descriptions of the abilities or functions and structures responsible for realizing them. A nonexhaustive list of such metatheoretical assumptions would include, for example: the representation of time and temporality as a fundamental organizing component of the apperceptive subject of experience, the objectivity of categories which may very well be subjective (as per Hegel's critique of Kant), the view of natural languages as unaffected by or free from psychologistic residues of representation, etc. The primary locus of these hidden metatheoretical assumptions are the categories by which we perform general classifications, giving structure to the world and our experience in it. Categories, as Kant rightly observed, are not the products of particular experiences or encounters with items in the world, but rather rule-governed invariances or general concepts organized by the manner, the *modus operandi*, by means of which the mind universally organizes sense-given materials. They are patterns of the mind's patterning of all that is sensed (abstractings), not patterns abstracted from what is sensible (abstracteds). Without abstracting qua act, there wouldn't be any abstracted.

In slightly more contemporary terminology, what Kant calls categories or pure concepts of the understanding are general classificatory functions capable of integrating local invariants synthesized from sense-given materials, and therefore of constructing rule-like generalities for the ordering and construction of objects: identification of local invariants (generic judgements), reidentification of local invariants in our different encounters with items in the world (recognitive judgements), and classification of local variations of particular items/objects (predictive judgements). But the problem with categories as necessary and universal forms of experience is that the extent to which they are bound up with the particular, local, and contingent aspects of these sense-given materials is far from obvious and is not yet a settled issue. However this faculty of understanding, and the experience of the world it has thus-and-so categorially formatted, can themselves be subjected to judgements that bring forth both the cracks and possible openings in our experience. The metatheoretical assumptions of the structuring categories precisely herald the as yet unknown extent to which what we perceive as universal and necessary structuring acts of cognition may in fact be particular, local, and contingent aspects of our experience.

Toy models come in small and big varieties. The small toy model is a simplified version of only one theoretical model (essentially it is a model in a collapsed form), whereas big toy models are models that accommodate different and often seemingly incompatible models and theories, such as general relativity and quantum mechanics. Put differently, big toy models represent a compressed form of model pluralism, and in order to do so they are required to have a conceptual architecture plastic enough to accommodate and faithfully represent the main features of different theoretical frameworks, while at the same time being capable of preserving the contrasting features of these accommodated systems as distinct categories. Here 'category' refers to the category-theoretical sense of mapping objects and their relationships. In order to for us to be able to adequately think about the kind of problems that we are dealing with when talking about the construction of AGI, we first need a big toy model. An AGI big toy model should be able to coherently accommodate different models derived from physics, evolutionary biology, neuroscience, developmental psychology, multi-agent system design, linguistics, logic, and computer science. One of the problems with old AI research was that it was strongly driven by unique and inflationary models of mind that generated more setbacks than steps forward. For example, consider the symbolic program of AI (the syntactic picture of the mind), deep learning (neural networks and statistical inference), and computational semantics (computational-logical modelling of meaning representation in natural languages), programs which were developed based on insights derived from the evolutionary sciences, computer science, neuroscience, logic, and linguistics. While in their own right these programs have led to undeniable achievements and progress in the field of artificial intelligence, they have also created theoretical bottlenecks and practical setbacks. This is because their implicit metatheoretical assumptions have been either left uncontested owing to the sheer success of their ideas and methods within a narrow domain of application, or unduly overstretched into global assumptions about the nature of cognition and mind. The result is that the statistical framework of something like machine learning becomes the global model of general intelligence, or the characteristics of sequential algorithms (effective mechanizability and symbol-manipulation) establish a syntactic model of mind within which the program of artificial intelligence as a whole is oriented. Once the metatheoretical assumptions of these locally successful ideas and methods are inflated into global models of general intelligence or mind, it is only a matter of time before the model arrives at theoretical and practical impasses, and development comes to a halt. The summer of AI, as we have known it thus far, turns out to be a long if not perpetual winter.

Toy models, on the other hand, are not only explicit metatheories in themselves, but also make explicit the implicit metatheoretical frameworks of their constituent ideas, observations, and methods. By doing this, toy models are able to keep these implicit metatheoretical assumptions underlying their theoretical commitments in check, and therefore avoid the risks of inflationary models. Their utility lies not only in the idea that they permit some theoretical arbitrage by combining and spanning different metatheories, but also, and more importantly, in their ability to facilitate the reinterpretation, reassessment, and reapplication of conventionally interpreted ideas and observations.

However the real value of a toy model is that one learns from it by breaking it in the real universe; but not until one has systematically played with it. It is exactly in this sense that a toy model or toy universe of AGI is an *explicit metatheory* of artificial general intelligence constructed from falsifiable concepts and models drawn from different theoretical frameworks