TTQ PAPER SUMMARY

Aaron Turner, 15 September 2026 (updated 26.9.26)

Because the full paper (a) addresses the ASI Endgame Problem (AEP), of which Normative ASI Superalignment (NAS) is just a sub-problem, and (b) attempts to do so extremely thoroughly, there is A LOT of stuff in the full paper. I will attempt to summarise:

1. Abstract — Top-Down Breadth-First Recursive Decomposition (TDBFRD) applied to the AEP generates a large number of sub-problems, most of which are only partially addressed within the paper (instead generating multiple sub-problems to be further addressed by the next TDBFRD iteration); NAS is the ONLY problem generated in this iteration that is addressed by the paper all the way to a substantive solution (albeit with a couple of loose ends to be addressed at implementation time); accordingly, the abstract (and paper title) focusses primarily on NAS and the paper's proposed TTQ+NAP solution.
2. Video — The paper opens by pointing the reader to the TL;DR video (which at the time of writing isn't yet finalised and so will eventually need updating with the final version).
3. Scope — Describes the TDBFRD approach to the AEP, and how the current paper corresponds to the first TDBFRD iteration, whereby each such iteration generates additional information (likened to a "layer of lasagne") contributing, probably in multiple ways, towards an AEP solution; being the "top" layer, the current paper is largely conceptual in nature.
4. Introduction —
    1. Describes our ultimate objective ("maximise the net benefit of Artificial General Intelligence (AGI) for all humanity, without favouring any subset thereof"), and the two interdependent tracks (AGI R&D and AGI governance) of which the problem is comprised;
    2. defines agentic superintelligent AGI ("fully autonomous systems (able to sense and affect the physical world) that are more intelligent than any human"), and the ASI Endgame ("the point in time at which humanity develops agentic superintelligent AGI"), which is prudently assumed to be an irreversible trapdoor.
    3. describes the spectrum of all possible ASI Endgames, and defines the AEP as "How can we ensure that the ASI Endgame that actually transpires is as close as possible (and ideally identical) to the ideal ASI Endgame (from the perspective of the human species as a whole)?".
    4. Introduces Gold-Standard (GS) AGI as AGI that is maximally-aligned ("impossible to be more strongly aligned with humans") and maximally-validated ("impossible to be more convincingly proved correct") [NOTE: for technical reasons, these are later redefined as practical-maximal-alignment and practical-maximal-validation]
    5. Argues that an ASI Endgame in which all deployed agentic superintelligent AGI satisfies the Gold-Standard would correspond to the ideal ASI Endgame (from the perspective of the human species as a whole); accordingly, the AEP decomposes into somehow achieving this state of affairs (via a coordinated combination of the AGI R&D and AGI governance tracks):
        1. AGI Governance — this track seeks to achieve \forall x: AGI(x) => GS(x) in two stages: (a) define (and then lobby for) an International Standard for Gold-Standard AGI based on the definitions of practical-maximal-alignment and practical-maximal-validation (later) defined in this paper, and (b) define (and then lobby for) a Gold-Standard AGI Treaty mandating (with strict enforcement and criminal penalties) that any AGI must be certified (by a competent independent certification authority) as satisfying the Gold-Standard prior to deployment; this is as far as the current paper takes the AGI Governance track, which will be taken up by a later paper
        2. AGI R&D — this track seeks to achieve \exists x: ASI(x) & GS(x), i.e. to build a Gold-Standard ASI (i.e. an agentic superintelligent AGI), which is then decomposed into a further two stages:
            1. normative alignment — how do we define a final goal FG which, if pursued as intended, ensures practical-maximal-alignment?
            2. technical alignment — how do we build a practically-maximally-validated agent G that pursues FG as intended?
        3. given that \forall x: ASI(x) => AGI(x), together these two tracks (if successful) will achieve (\forall x: AGI(x) => GS(x)) & (\forall x: ASI(x) => GS(x)) & (\exists x: ASI(x) & GS(x))
        4. the remainder of the current paper focusses (almost entirely) on normative alignment, and technical alignment is (almost entirely) deferred to a later paper
    6. Contains an outline of the current paper, including a recommended engagement funnel:
        1. watch the TL;DR video
        2. read sections 2, 3, and 13
        3. read section 12
        4. carefully (re-)read sections 2-13.

7. Contains (a) a glossary of terms, and (b) a list of "quick links" pertaining to well-known alignment failure modes etc
8. States that the Gold-Standard ASI implemented by the AGI R&D track (a) will be called BigMother, (b) will be owned by all of humanity (e.g. via the UN) as a global public good, and (c) will be operated in such a way that benefits all of humanity, without favouring any subset thereof (exactly as per the ultimate objective stated earlier).

5. Concepts and definitions pertaining to AGI —
    1. defines intelligence (as problem solving), AI, problem-solvers, and agents
    2. defines problem statements (e.g. expressed as a triple <P, Q, R>), and problem solutions
    3. describes how a problem-solver strives to find valid solutions by searching the space of all possible solutions
    4. argues that problem-solving (and therefore intelligence) has three scalable dimensions:
        1. inventiveness (ability to use information to reduce the search space)
        2. knowledge (i.e. the information used to reduce the search space)
        3. physical resources (e.g. time, energy, compute)
    5. defines artificial narrow intelligence (ANI), and artificial general intelligence (AGI), both in terms of problem solving
    6. defines taxonomies for ANI (1-3) and AGI (1-7)
    7. defines superintelligent ANI, superintelligent AGI, and maximally-superintelligent ANI and AGI
    8. gives some examples of real-world AI systems; note that all the current frontier models are classified as AGI 4 under my taxonomy (and I suspect that they will never break out of AGI 4)
    9. describes the (logical) internal architecture of an agent
    10. discusses extrinsic (externally observable) intelligence vs intrinsic intelligence (the latter of which requires an (ideally sophisticated) internal world model — i.e. theory of the physical universe ToPU — that captures the semantics of the problem domain, such as how the universe operates causally)
    11. describes Chinese Rooms as having non-trivial extrinsic intelligence but zero (or at most trivial) intrinsic intelligence (given the available mech interp etc evidence, I suspect that the current frontier LLM-based chatbots are all Chinese rooms, with significant extrinsic intelligence (and therefore utility, which also makes them potentially dangerous) but trivial intrinsic intelligence)
    12. introduces the alignment and superalignment problems (in terms of normative and technical (super)alignment)
6. Some philosophical issues pertaining to AGI
    1. the longest section in the paper (~ 75 pages)
    2. defines classical computation and cognitive computation
    3. argues that (at the very highest level of abstraction) human cognition is (a) a cognitive computation, and (b) (substantially) less than perfect
    4. argues that superintelligence is both possible and (almost certainly) inexorable
    5. defines (for our purposes) the mathematical universe (and theories thereof), the physical universe (and theories thereof, including theories of causality, mind, and language), and a top-level theory of the universe (i.e. belief system) encompassing all information in the system (including the theories-of-the-mathematical-and-physical-universes)
    6. briefly discusses consciousness and its relevance to AGI (this becomes very important later)
    7. building internal world models through continuous learning:
        1. outlines (in ~ 40 pages) a continuous learning algorithm (as discussed in point 18 of the AGI Research History doc) called UATL that takes an agent's percept history as input and constructs as output a sophisticated theory-of-the-physical-universe ToPU represented as a fully-interpretable, explicit data structure, with a known semantics
        2. starts with the LLM learning objective (for predicting the next token)
        3. derives an agent learning objective (for predicting future percepts)
        4. argues that predicting future percepts is a strong enough training signal from which to construct a sophisticated theory-of-the-physical-universe ToPU, whereas simply predicting next tokens as per LLMs is (almost certainly) not
        5. concludes that (given current evidence, e.g. from mechanistic interpretability) LLMs are essentially mere Chinese Rooms, lacking non-trivial internal world models (note: OthelloGPT is a trivial world model, orders of magnitude less sophisticated than those potentially constructed by UATL) and therefore lacking non-trivial intrinsic intelligence
        6. argues AGAINST implementing UATL (or any major agent component of an AGI/ASI) as an end-to-end NN, for 3 reasons:
            1. alignment — as discussed in point 23 of the Research History doc, (given our current understanding) NNs cannot be controlled with sufficient engineering precision for the purposes of reliable alignment, let alone superalignment — this is also the primary conclusion of Yudkowsky and Soares's "If Anyone Builds It, Everyone Dies", which is cited
            2. validation — in order to satisfy the (not at this point formally defined) validation requirement for Gold-Standard AGI, we need to apply formal

verification to all hardware and software components, which (because they are not yet sufficiently well understood) cannot be applied to NNs
3. interpretability — in order to be of any use to any other agent component (and also for validation/verification purposes) the resulting theory-of-the-physical-universe ToPU must be fully interpretable

7. starting with Plato's Allegory of the Cave, describes a two part learning process (Up and then Left):
    1. Up — a perceptual pipeline (from physical percepts to logical percepts to compound percepts) that first captures as much perceptual informationas possible (as compound percepts ~= perceptual events), and then constructs (via global induction over compound percepts) a Perceptual Concept Hierarchy (PCH), effectively comprising the agent's (subjective) ontology of perceptual concepts (ranging from raw perceptual events at the bottom, to "thing" at the root of the hierarchy)
    2. Left — having captured every observed (perceptual) pattern, pattern of patterns etc in the PCH, the UATL algorithm then hypothesises (via abduction) a theory-of-the-physical-universe in 2 phases:
        1. the canonical ToPU is a simple "mirror" of the patterns in the PCH; that is, every observed perceptual feature in PCH becomes a (hypothesised) physical feature in the canonical ToPU; this is broadly equivalent to a human 4-year-old's (behaviourist) internal world model, and does not include any physical features (e.g. Neptune) that the agent hasn't directly observed
        2. a set of candidate ToPU^js is then derived from the canonical ToPU via the scientific method (basically, an information-driven search of the space of all possible ToPU^js for new ToPU^js that out-perform all other ToPU^js), where each new candidate ToPU^j is measured for its ability to predict future percepts (the ultimate "ground truth"), and the best performing ToPU^j is chosen as "current best"
8. discusses active learning (where the continuous learning mechanism is able to instantiate new sibling goals for the continuous planning mechanism such as "what's under that rock?")
9. discusses metacognition (global induction over the history of its own PCHs, ToPU^js, etc, looking for patterns — this is how an agent can learn epistemic humility (by itself, via introspection), which is later incorporated into an NAP requirement)
10. as a corollary, also briefly discusses metalearning and metaplanning
11. discusses a number of ways in which the UATL (and similar) algorithms may be accelerated — in particular, "System 1 + System 2 = neurosymbolic" outlines how NNs can be incorporated into an agent's lower-level implementation while still satisfying the formal verification requirement of Gold-Standard AGI

8. Finally, discusses a number of "big questions", including "what is understanding?", for which (having gone through the high-level development of UATL) the paper is now able to give a computational answer — this computational definition of "understanding" is later referred to several times in the NAP requirements (e.g. S-minus must understand X)

7. Communication between two cognitive computations
    1. because we will eventually be COMMUNICATING the final goal TTQ to S-minus as a MESSAGE (rather than trying to grow S-minus as a trained NN with TTQ somehow "baked in"), the paper considers the process of communicating a message (such as TTQ) between two intelligent entities (such as S-minus's designers and S-minus)
    2. the paper tries to consider all the possible communication failure modes
    3. given that so many things can go wrong, the paper asks "how is communication even possible?"; answer:
        1. Alice and Bob both (i.e. sender and recipient) share the same physical universe
        2. Alice and Bob are both intelligent (and therefore able to at least partially recover Alice's intended meaning within Bob)
    4. the paper considers how communication errors may be mitigated through intelligent action
    5. the paper also considers a number of critical/important decision points, e.g.:
        1. choice of abstract encoding method (i.e. message language/notation), e.g.:
            1. pure maths
            2. natural language (such as English)
            3. applied maths (mathematical model)
        2. how the abstract message is formulated

8. Concepts and definitions pertaining to alignment
    1. safety (bad things don't happen) and liveness (good things do happen)
    2. alignment = liveness + safety (good things happen, and bad things don't — a key meme in the safety-critical world introduced by Lamport in 1977)

3. human values, and how they might be determined — this is where the paper concludes that a fixed set of predetermined values would be inappropriate, and that the agent must instead work at a meta level (i.e. observing humans, and trying to work out their values/preferences from these observations); I recognised the broad similarity with Russell's CIRL, and cited Russell's work in a footnote
4. ordinal preferences, utility functions, expected utility
5. eliciting (idealised) human preferences
    1. very much with the 2016 Brexit referendum in mind, I started by considering the various ways in which preference elicitation might go wrong (e.g. some people are less than entirely rational, or less than entirely well-informed, some have been fed mis/disinformation (and believed it), and some have been purposefully manipulated); from this, I reasoned that "(human) H's best interests would be much better served if agent G were to carefully observe H and, from those observations, estimate (as accurately as possible) what H's actual rational well-informed freely-determined preferences (expressed as e.g. an estimated utility function U) would be if H were maximally (a) rational, (b) well-informed, and (c) free from manipulation by any party", which then became the exact wording I used in TTQ in respect of idealised preferences; I recognised the broad similarity with CEV, which I cited in a footnote; on 10 October 2024 I emailed Stuart Russell, Roman Yampolskiy, and Eliezer Yudkowsky with snippets taken from the current version of the paper (which I believe was then called "The BigMother Manifesto - Part 1"); Stuart Russell replied "seems quite similar to CIRL (aka assistance games) and to Yudkowsky's CEV", Roman Yampolskiy replied "how is "strive to estimate (as accurately as possible) what B's actual rational well-informed freely-determined (i.e. idealised) preferences P would be if B were entirely rational, well-informed, and free from manipulation by any party" different from *"if we knew more, thought faster, were more the people we wished we were, had grown up farther together"?*" (i.e. CEV), and Eliezer Yudkowsky replied "What is this supposed to do that Coherent Extrapolated Volition does not do?"; I'm afraid that my replies at the time were very poor — being more of an engineer than an academic, I hadn't yet worked out how to articulate what was academically new
    2. I also indicated how H's idealised preferences might be calculated by an agent G that maintained a ToPU c/o UATL — G can formulate a modified version of ToPU (the counterfactual) and then deduce (i.e. estimate, via deduction) from it H's idealised preferences; this is only possible if ToPU is represented logically, e.g. in FOL/FOT
6. aggregating human preferences (maximally fairly)
    1. perfect alignment is impossible (Arrow's theorem, etc)
    2. the paper outlines a proposed mechanism for maximally-fair aggregation (basically an existence proof for possible solutions), and suggests that the same approach might be applied to more complicated examples such as population gaming
7. realising human preferences — conceptually straightforward, the paper suggests maximising expected utility as one possibility
8. statistical definitions (in terms of discrete probability functions) of maximal-alignment (one of the primary elements of Gold-Standard AGI), robust-alignment, indifferent-alignment, and maximal-misalignment
9. careful discussions of what can happen if an agent is less than robustly aligned: existential risk, the risk spectrum
10. the Global Corporate Capitalist Market Machine (GCCMM) as an example of a misaligned superintelligence (which also happens to be driving the current furious race to AGI/ASI)
11. examples of various global shocks (which might be triggered by a misaligned AGI) and their recovery times – all of this gloom and doom is intended to induce a safety-critical mindset

9. Concepts and definitions pertaining to validation
    1. reviews:
        1. the system development pipeline
        2. system development errors
        3. validation, verification, and formal verification
    2. argues that validation effort must be proportionate to downside risk (and that AGI >= 3 is not just safety-critical, but existentially-critical)
    3. reviews formal verification, and its limitations
    4. reviews a toolbox of validation techniques (which may be used to catch various errors that formal verification cannot)
    5. defines maximal-validation (the other primary element of Gold-Standard AGI)
10. Gold-Standard AGI in actual engineering practice
    1. because maximal-alignment and maximal-validation are asymptotes (unattainable in practice, because they require infinite information, compute, and/or time), this section defines:
        1. practical-maximal-alignment (PMA) — the objective that we are trying to achieve with TTQ+NAP

        2. practical-maximal-validation (PMV) — uses standard safety-critical engineering to determine a validation stopping point
    2. reviews safety-critical terminology, hazards, tolerable risks, risk reduction, safety cases
    3. estimates (by comparison with e.g. local supernovae) the societal tolerability of the AGI-assisted termination of humanity (as once every 10^9 - 10^12 years) — this then becomes the target MTBF (Mean Time Between Failure) for any deployed superintelligent agent
    4. safety engineering applied to Gold-Standard AGI (which is not a “normal” safety-critical system)
11. Formulating a mandate for a guardian — describes the babysitter analogy, and 5-phase process
12. CASE 1 — as an example, goes through the 5 phases in the case of hiring a babysitter
13. CASE 2 — goes through the exact same 5 phases in the case of designing and building a superintelligent agent
    1. this is the main part of the paper, i.e. the part where the proposed TTQ+NAP solution is formulated
    2. this section refers back to every other section in the paper — i.e. this is where everything finally comes together
    3. quick review of things that we do and do not want to happen (when S is ultimately deployed)
    4. Phase 1 — formulate the final goal FG; we follow the advice of the earlier “Communication” section:
        1. Choice of abstract encoding method (message language/notation) — counter-intuitively, we conclude that, due to the fact that the final goal for an agent must necessarily refer to the physical universe, FG must be expressed in natural language, for which we choose 2YYY CE (the year the final goal is finalised) English
        2. Formulation of the abstract message — this is where we formulate FG (in English)
            1. minimise ambiguity
            2. no pressure! — because S is superintelligent, we cannot safely rely on any monitoring / oversight / control mechanism other than FG as formulated here
            3. use of structured English (with braces) to minimise ambiguity
            4. careful definition of tasks T1 and T2 (liveness); analysis of consequent properties etc
            5. careful definition of qualification Q (safety); analysis of consequent properties
            6. careful definition of the assumptions part, with explanation
            7. careful definition of the imperative part; with explanation
            8. putting it all together as TTQ
    5. Phase 2 — identify sufficient conditions (NAP)
        1. we now work BACKWARDS from TTQ in order to determine NAP such that NAP-compliance combined with the wording of TTQ implies that superintelligent agent S will be practically-maximally-aligned (PMA)
        2. note that this is NOT simply a definition of superintelligence; rather, it is DERIVED from TTQ; for example:
            1. we know that TTQ is in English, so S-minus must be able to understand English
            2. we know in particular that S-minus will need to understand certain words and phrases in TTQ
            3. we know that S-minus must be motivated to pursue TTQ
            4. we know that pursuance of TTQ will involve generic problem-solving
            5. we know that generic problem-solving requires:
                1. rationality (epistemically justified beliefs, valid logical reasoning etc)
                2. information relevant to the problem domain
                3. physical resources (e.g. compute)
            6. we know that, in order to be robustly-aligned, S-minus must have at least a minimum understanding of what it means to be safely pro-human
        3. in order to be maximally confident that TTQ+NAP => PMA (which is our ultimate design objective for superintelligent agent S), we take the basic NAP requirements and then generalise them to the maximum extent possible while still being finite and therefore physically implementable (and validatable) — the result is NAP proper
        4. note that 3 of the NAP requirements have the form “S-minus must have at least as good an understanding of X as any baseline human”, where “understanding” is as defined (computationally) in the philosophy section
        5. note that one of the requirements (well-resourced) mandates that S-minus’s HW must not be phenomenally conscious (i.e. sentient), which was strongly recommended in the philosophy/consciousness section (for both safety and ethical reasons)
    6. Phase 3 — design, build, and validate S-minus
        1. describes a procedure for designing, building, and validating S-minus (as per PMV)
        2. this explicitly includes a validation step which asks if PMA is in fact what humans want
        3. assumes the existence of a competent independent certification authority (one of the objectives of the governance track) that examines the validation case (against the Gold-Standard) and determines whether or not S can licensed for deployment

7. Phase 4 — transmit the final goal FG — (in the example used in the paper) plug TTQ (which the paper assumes has been blown into a PROM) into S-minus, thereby giving a fully operational, fully autonomous superintelligent agent S
8. Phase 5 — superintelligent agent deployment; S could potentially faithfully serve humanity for millions of years (assuming of course that NAP-compliance, together with the wording of TTQ, implies PMA)
9. Evaluation
    1. does NAP-compliance (together with TTQ) imply PMA? — the paper provides a careful but nevertheless informal argument in the affirmative — THIS IS THE PRIMARY RESULT OF THE PAPER
    2. the TTQ challenge (asks readers to try and break TTQ)
    3. possible failure modes — examines some corner cases in which TTQ+NAP might fail, either logically, or physically, and how/why these may be mitigated/tolerated
    4. argues that S will always execute TTQ (perhaps not perfectly, but) at least as well as any baseline human, or any collection of baseline humans
    5. fundamentally, this is because we leveraged S's superintelligence when formulating TTQ+NAP
    6. can we design and build an NAP-compliant S-minus?
        1. strictly speaking, this is outside the scope of the current paper
        2. nevertheless, the paper outlines an 8-step construction sequence for S-minus (to be expanded upon in a later paper on technical alignment)
    7. the NAP challenge (asks readers to try to design/build an NAP-compliant S-minus)
    8. formalisation — indicates how, at implementation time, the earlier informal argument that NAP-compliance (plus TTQ) implies PMA might be formalised, potentially up to and including fully formal proof
    9. refinement
        1. potential tweaks to NAP (so it's easier to implement, but still implies PMA)
        2. argues against potential tweaks to TTQ in respect of so-called "red lines"

14. Conclusion:
    1. contributions to the AEP (in the order in which they appear)
    2. limitations
    3. methods
    4. related work — focussing on prior art, and novelty
    5. further work, i.e. currently planned future TDBFRD iterations