On Eudaimonia and Optimization

I.

What follows for AI alignment if we take the concept of eudaimonia -- active, rational human flourishing -- seriously? I argue that the concept of eudaimonia doesn't simply point to a desired state or trajectory of the world that we should set as an AI's optimization target, but also to a form of practical rationality different from standard consequentialist[1] rationality. I propose that this form of rational activity and valuing, which l call *eudaimonic rationality* or *praxis*, may prove useful or even necessary as a framework for the agency and values of AIs themselves. I give several arguments to the effect that if we wish to align AIs with human flourishing, we best direct AIs towards practicing a form of agency closely related to the form of agency involved in human flourishing itself. The arguments are based both on the dangers of a 'type mismatch' between human flourishing as an optimization target and consequentialist optimization as a form, and on certain material advantages that eudaimonic rationality plausibly possesses in comparison to deontological and consequentialist agency with regard to stability and safety.

The concept of eudaimonia, I argue, suggests a form of rational activity without a strict distinction between means and ends, or between 'instrumental' and 'terminal' values. In this model of rational activity, a rational action is an element of a valued practice in roughly the sense

[1] Recall that because of the possibility of 'notational consequentializing' (rewriting any policy as a utility function), dividing agents or even theories or decision-procedures into 'consequentialist' and 'non-consequentialist' isn't a strict formal distinction. Throughout this essay, by 'consequentialist' I will mean roughly an agent for whom, in ideal practical reasoning, means and outcomes are effectively separately evaluable and the value of outcomes is typically decisive. Semi-formally, by 'consequentialist' I mean an agent *s* such that when *s* considers whether to perform action *c*, *s*'s ideal reasoning is an expected-utility calculation using a utility-function whose utility-assignment to a world-trajectory *w* has *low sensitivity* to whether *s* performs *c* in *w* (holding everything else about *w* constant).

in which a note is an element of a melody, a time-step is an element of a computation, and a moment in an organism's cellular life is an element of that organism's self-subsistence and self-development.[2] This is by contrast with a typical means-end understanding of rational activity, where a rational action is a means to a valued end logically independent from the action.

Crucially, my argument that AI alignment should aim to instill AIs with a eudaimonic form of agency (or at least take close formal inspiration from eudaimonic rationality) is not primarily an ethical argument. The core of my argument is rather that our best intuitions about the nature of human flourishing are implicitly intuitions that what I call *eudaimonic rationality* can -- with the right content -- be functionally robust in a sense highly relevant to AI alignment. More specifically, I argue that in light of our best intuitions about the nature of human flourishing it's plausible that eudaimonic rationality is a *natural* form of agency, and that eudaimonic rationality is *efficacious* even by the light of certain consequentialist approximations of its values. I then argue that if our goal is to align AI in support of human flourishing, and if it is furthermore plausible that eudaimonic rationality is natural and efficacious, then many classical AI safety considerations and 'paradoxes' of AI alignment speak in favor of trying to instill AIs with eudaimonic rationality or with decision-procedures closely inspired by eudaimonic rationality.

Throughout this essay, I will sometimes-explicitly and often implicitly be asking whether a certain form of agency or rationality or practice is *natural*. The sense of 'natural' I'm calling on is certainly related to the senses used in various virtue-ethical traditions, but the interest I take in it less immediately normative and more material or technical. While I have no reductive definition at hand, the intended meaning of 'natural' is related to stability, coherence, relative

[2] This way of thinking is broadly associated with analytic neo-Aristotelians such as Alasdair MacIntyre and Michael Thompson.

non-contingency, ease of learnability, lower algorithmic complexity, convergent cultural evolution, hypothetical convergent cultural evolution across different hypothetical rational animal species, potential convergent evolution between humans and neural-network based AI, and targetability by ML training processes. While I will also be making more direct reference to AI alignment, it is really this material concept of naturalness that is doing the work: if there is reason to believe that certain exotic-sounding forms of agency, rationality, or practice are both themselves natural and make our moral and eudaimonic values natural in turn, then we have learned about good, relatively safe, and relatively easy targets for AI alignment.

II.

I start with a consideration of the nature of the good we hope AI alignment can promote. With the exception of hedonistic utilitrians, most actors interested in AI alignment understand our goal as a future brimming with human -- and other sapient-being -- flourishing: persons living good lives and forming good communities. What I believe many fail to reflect on, however, is that on any plausible conception human flourishing involves a kind of *rational activity* as a constitutive part of the activities through which we flourish. Subjects engaged in human flourishing act in intelligible ways subject to reason, reflection, and revision, and this form of rational care and purposefulness is itself part of the constitution of our flourishing. I believe this characterization of human flourishing is relatively uncontroversial upon reflection, and yet it raises a kind of puzzle if we are used to thinking of rationality in consequentialist (or consequentialist-with-deontological-constraints) terms: Just what goal is the rational agency involved in human-flourishing activity directed towards?

One obvious answer would be that, like all properly aligned rationality, the rational agency involved in human-flourishing activities is geared towards maximizing human (and other sapient) flourishing. I think reflection quickly finds, however, that we must be confused or conflicted about the right way to describe the contribution that rational agency in human-flourishing activities makes to human flourishing: It seems neither appropriate to say that the rational agency involved in a human-flourishing activity contributes to human flourishing simply by *constituting* an instance of human flourishing (by selecting actions that are intrinsically valuable when rationally selected), nor appropriate to say that the rational agency involved in a human-flourishing activity contributes to human flourishing just *instrumentally* (by selecting actions that causally promote human flourishing).[3]

Since 'human flourishing' can seem mysterious and abstract, let us focus our discussion here by narrowing our scope to some concrete eudaimonic practices.[4] Consider practices like like math, art, craft, friendship, athletics, romance, play, and technology, which are among our best-understood candidates for partial answers to the question 'what would flourishing people in a flourishing community be doing.' From a consequentialist point of view, these practices are all marked by extreme ambiguity -- and I would argue indeterminacy -- about what's instrumental and what's terminal in their guiding ideas of value. Here, for example, is Terry Tao's account of goodness in mathematics:

> 'The very best examples of good mathematics do not merely fulfil one or more of the criteria of mathematical quality listed at the beginning of this article, but are more importantly part of a greater mathematical story, which then unfurls to generate many

[3] Technically speaking the first horn of the dilemma can be further bifurcated into 'rational agency contributes to human flourishing by choosing actions that are intrinsically valuable however chosen' and 'rational agency contributes to human flourishing by selecting actions such the these actions combined with their selection by rational agency are intrinsically valuable.'

[4] I would suggest that practices like math, art, craft, friendship, athletics, romance, play, and technology, aside from almost certainly belonging in the tapestry of human-flourishing activity, are also in themselves of the right logical type for 'flourishing': a mathematical field (or a person's mathematical life) can wither or flourish, a friendship can wither or flourish, technological development can wither and flourish, and so on.

> further pieces of good mathematics of many different types. Indeed, one can view the history of entire fields of mathematics as being primarily generated by a handful of these great stories, their evolution through time, and their interaction with each other. I would thus conclude that good mathematics [...] also depends on the more "global" question of how it fits in with other pieces of good mathematics, either by building upon earlier achievements or encouraging the development of future breakthroughs. [There seems] to be some undefinable sense that a certain piece of mathematics is "on to something", that it is a piece of a larger puzzle waiting to be explored further.'

It may be possible to give some post-hoc decomposition of Tao's account into two logically distinct component -- a description of a utility function over mathematical achievements and an empirical theory about causal relations between mathematical achievements -- but I believe this would be artificial and misleading. On a more natural reading, what Tao is doing is describing are some of the conditions that makes good mathematical practice a *eudaimonically rational practice*: in a mathematical practice guided by a cultivated 'undefinable sense that a certain piece of mathematics is "on to something,"' present excellent performance by the standards of this practical-wisdom judgment reliably develops the conditions for future excellent performance by the standard of this practical-wisdom judgment, a well as cultivating our practical and theoretical grasp of the standard itself.[5]

Crucially, the mathematical practical-wisdom described by Tao is not fully conceptually opaque beyond its basic characterization as a self-cultivating criterion for self-cultivating excellence in mathematical activity. Mathematical flourishing can *partly* be described as involving the instantiation of a relation (a mathematical-practice relation of 'developmental connectedness') among instantiations of relatively individually definable and quantifiable instances of mathematical value such as elegant proofs, clear expositions, strong theorems and so

[5] See Tao: '[The] determination of what would constitute good mathematics for a field can and should depend highly on the state of the field itself. It should also be a determination which is continually updated and debated, both within a field and by external observers to that field; as mentioned earlier, it is quite possible for a consensus on how a field should progress to lead to imbalances within that field, if they are not detected and corrected in time. '

on. Furthermore, this relation of developmental connectedness is not just an intrinsically valued relation but one whose traits include a reliable tendency to causally propagate instances of relatively individually definable and quantifiable mathematical value (instances of elegant proofs, clear exposition, strong theorems and so on):

> [I believe] that good mathematics is more than simply the process of solving problems, building theories, and making arguments shorter, stronger, clearer, more elegant, or more rigorous, though these are of course all admirable goals; while achieving all of these tasks (and debating which ones should have higher priority within any given field), we should also be aware of any possible larger context that one's results could be placed in, as this may well lead to the greatest long-term benefit for the result, for the field, and for mathematics as a whole.

One could, again, try to interpret this causal relationship between excellence according to Tao's 'organicist' (or 'narrative' or 'developmental') sense of good mathematics and the reliable propagation of narrow instances of good mathematics as evidence of a means-ends rational relation, where additive maximization of narrow instances of mathematical value is the utility function and 'organicist' mathematical insight is the means. For Tao, however, the evidential import of this causal relationship is exactly the opposite -- this causal relationship suggests the *unification* of our myriad more explicit and more stand-alone conceptions of excellence in mathematical activity through a more ineffable but more complete conception:

> It may seem from the above discussion that the problem of evaluating mathematical quality, while important, is a hopelessly complicated one, especially since many good mathematical achievements may score highly on some of the qualities listed above but not on other. [However, there is the remarkable] phenomenon that good mathematics in one of the above senses tends to beget more good mathematics in many of the other senses as well, leading to the tentative conjecture that perhaps there is, after all, a universal notion of good quality mathematics, and all the specific metrics listed above represent different routes to uncover new mathematics, or different stages or aspects of the evolution of a mathematical story.

Tao's reasoning in the above points to what is, in my view, the core difference between consequentialist rationality and eudaimonic rationality when we regard them not only as procedures for making decisions in light of settled values but as paradigms for second-order rationality or 'reflective equilibrium' processes, where the determination of rational action involves engagement with theoretical questions about values. Within the paradigm of consequentialist rationality, if excellence[6] in accordance with a holistic, difficult-to-judge prima facie value (say 'freedom') is reliably a powerful causal promoter of excellence in accordance with more explicit, more standalone prima facie values (say 'material comfort,' 'psychological health,' 'lifespan'), this relationship functions as evidence *against* the status of the holistic prima-facie value as a constitutive -- as opposed to instrumental -- value. Within the paradigm of eudaimonic rationality, by contrast, this same relationship functions as evidence *for* the status of the holistic prima-facie value as a genuine value.

For a (typical)[7] *consequentialist-rationality reflective equilibration* process, evidence of the instrumental value of the excellence of a whole for the prima facie independently-valuable excellence of its parts *explains away* our investment in the excellence of the whole. The "coincidence" that the intrinsically valuable whole is also instrumentally valuable is taken to suggest a kind of double-counting error -- one we can "fix" without changing widely accepted action-prescriptions too much by concluding that the whole has no intrinsic value but valuing the whole is an optimal heuristic under normal circumstances. A *eudaimonic-rationality reflective*

[6] Within this essay I use 'excellence' as the most general, pre-theoretical term for accordance with a holistic evaluative standard. The standard can be instrumental or terminal, apply to actions or persons or states or objects, be moral or aesthetic or epistemic and so on, and the standard itself (and so the excellence it defines) may later be judged as rational or irrational, substantive or trivial, significance or insignificant.

[7] The above observation does not describe a formal feature of 'consequentialism' in accordance with some standard technical definition. However I believe it accurately describes a strong observable tendency in both the academic and 'rationalist' literature when conducting normative reflective equilibration within a consequentialist paradigm.

*equilibratio*n, by contrast, treats instrumental causal connections between excellences as evidence that our notions of excellence are picking out something appropriately 'substantive.'

For eudaimonic-rationality reflective equilibration, it is the discovery of causal and common-cause relations among excellences that ratifies our prima facie sense that caring about these excellence is eudaimonically rational. The discovery of these causal connections functions as evidence that:

> *1)* The 'local' excellences we care about are resonant or fruitful, in the sense that they are positively promotional of each-other and of holistic excellences in which they participate.
>
> *2)* The 'holistic' excellences we care about are materially efficacious and robust, in that the holistic excellences promote both the more local excellences that participate in them and their own continuation as future holistic excellence.

I importantly put 'local' and 'holistic' in scare-quotes in the above, since the relation of parts and wholes is likely iterable: Arithmetic geometry is part of algebraic geometry, which is part of mathematics, which is part of the arts and sciences, which is part of human culture, which is part of human flourishing, which may itself be part of other wholes to which the idea of excellence is applicable. Similarly, a practice capable of excellence may be part of multiple different wholes.

Before summing up my discussion of Tao's account of good mathematics, I'd like to risk a small digression to point out a potential strength of the eudaimonic-rationality approach: Eudaimonic rationality dissolves the 'paradox' that in real life arguments about the value of various human enterprises (e.g. the value of branches of science, branches of art, branches of sport) judgments of intrinsic value typically seek succor from some kind of claim to instrumental value. For example, a defense of the importance of research in quantum physics will often appeal to all the wonderful technological, mathematical, and special-sciences applications quantum

physics gave us, without meaning to reduce the worth of quantum physics to these applications. On my reading, these appeals aren't just additive -- 'aside from the intrinsic value there is also instrumental value' -- but presentations of evidence that research in quantum physics is a resonant part of a flourishing organic whole (e.g. the civilizational whole of 'modern science and technology'). I believe that without 'organicism' of this kind one faces a bad dilemma whenever one argues for the intrinsic worth of a pursuit or norm: either one stresses the value's independence from all benefits and applications and makes the claim of value dogmatic and irrelevant, or else one invites an instrumentalist reduction that 'explain away' the appearance of intrinsic value.[8]

Returning now to Tao's account of good mathematics, let's take stock of what my interpretation posited so far. I argue that mathematical excellence (the property marking 'the very best examples of good mathematics') according to Tao satisfies the following conditions, which I believe Tao intends as necessary but not sufficient:

> *A)* Mathematical excellence is a property of mathematical-activity instances.
>
> *B)* An excellent mathematical-activity instance performed today is excellent partly by virtue of satisfying the mathematical-practice relation '*builds on*' with regard to past excellent mathematical-activity instances.
>
> *C)* An excellent mathematical-activity instance performed today is excellent partly by virtue of having a reliable causal tendency to bring about future excellent mathematical-activity instances that satisfy the mathematical-practice relation '*builds on*' with regard to it.

[8] It may be fruitful to explore the potential of PageRank-like algorithms as theoretical models of how eudaimonic reflective equilibration works, and especially of how initial ideas of eudaimonic excellences are 'bootstrapped' from simpler and more local prima facie goods (and prima facie ills) in the first place. Scott Aaronson and Simon Dedeo have both discussed conceptual applications of PageRank-like algorithms in philosophy in various informal contexts. That said, I believe it's unlikely that PageRank over reliable instrumental-contribution relationships among prima facie goods and ills is the *full story* about the emergence of intrinsically valued holistic excellences, since while organicist relations between the excellence of wholes and of parts do involve instrumental-contribution relationships they plausibly also involve more rarified, 'hermeneutic' relations of (e.g.) mutually dependent intelligibility.

D) Instantiation of more local, more individually measurable criteria of mathematical-activity goodness such as elegant proofs, clear expositions, and strong theorems is a typical correlate of mathematical excellence.

E) At a given moment in a given mathematical field, the instantiation of mathematical excellence will be predictably better-correlated with the instantiation of certain local criteria of mathematical-activity goodness than with others.

Should we take these traits to collectively describe something more like a decision-procedure called 'mathematical excellence' that mathematicians should try to follow, or something more like an event called 'mathematical excellence' whose aggregate future-occurrences mathematicians should aspire to maximize? My contention is that Tao's account is inherently ambiguous, and for a good reason: there is no significant practical difference between **doing excellent mathematics** and **doing instrumentally optimal mathematics with regard to maximizing future aggregate excellent mathematics**. This isn't to say that doing excellent mathematics is the instrumentally optimal action **among all possible actions** with regard to aggregate future excellent mathematics, but that it is the (at least roughly) instrumentally optimal choice from ***among mathematical actions***.

I propose that the rough matchup between mathematical excellence and optimal (among mathematical actions) instrumental promotion of aggregate mathematical excellence is neither an empirical miracle nor something determined 'by definition' in a trivializing sense. Rather, 'mathematical excellence' as used by Tao is a concept that has a referent only if there is a possible property *x* that satisfies both desiderata *A*-*E* and the additional criterion that among mathematical actions, actions that are optimal as instantiations of *x* are also roughly optimal for maximizing aggregate future instantiation of *x*-ness.[9]

[9] We should perhaps also be concerned with the possibility of failures of uniqueness, as well as failures of existence.

This is what I would describe as the *material efficacy* condition on eudaimonic rationality. In order for a practice to be fit for possessing internal criteria of flourishing, excellence, and eudaimonic rationality, a practice must materially allow for an (at least roughly) optimally self-promoting property *x* that strongly correlates with a plethora of more local, more individually measurable properties whose instantiation is prima facie valuable. Stated more informally, there must exist a two-way causal relationship between a practice's excellence and the material, psychological, and epistemic effects of its excellence, such that present excellence reliably materially, psychologically, and epistemically promotes future excellence.

III.

I earlier said that if human flourishing is a practice of eudaimonic rationality, there may well be a kind of 'type mismatch' between human flourishing and the kind of consequentialist optimization we often associate with the idea of a future, agentially mature AI. In fact, I believe that implicitly recognizing but misdiagnosing this type mismatch is at least partially responsible for MIRI-style pessimism about the probability of aligning any artificial agents to human values.

On my view, the secret to relatively successful alignment among humans themselves (when there is successful alignment among humans) lies in the role *attempted excellence* plays as a filter on human interventions in the future trajectory of a eudaimonic practice. To the degree that humans value a given eudaimonic practice, they are committed to effecting their vision for the practice's future-trajectory primarily by attempting acts of excellence in the present, staking their intended effect on the practice's future-trajectory on the intended self-propagating excellence of their action. excellence of their acts. While this 'filter' doesn't necessarily stop the worst interventions from being harmful (there are forms of 'anti-excellence' that also have

certain self-promotion powers), I contend that it is mechanically crucial for the very possibility of reliably benign or positive interventions.

What do I mean? Consider the difference between a world where scientists typically try to propagate (what they believe to be) the scientific truth mainly by means of submitting research work to scientific institutions, and a world where scientists typically try to propagate (what they believe to be) the scientific truth by means including propaganda, fraud, threats, bribery, and slander. As Liam Kofi Bright demonstrates in On Fraud, a community of consequentialist scientists devoted to maximizing truth will predictably match the latter model. I believe one lesson to be drawn is that humans' ability to collaborate in the promotion of science depends on our ability to *scientifically* collaborate in the promotion of science, rather than throttle the future trajectory of science every-which-way our financial and political powers based on our individual beliefs about the optimal trajectory of science.

A flourishing eudaimonic practice is, above all, a natural-selection-like mechanism whose fitness-function selects among attempted acts of excellence the ones conductive to (and constitutive of) the practice's flourishing, propagating the excellence these acts instantiate. When people committed to a eudaimonic practice conduct their attempted interventions in the future trajectory of the practice via acts of attempted excellence, the natural-selection-like mechanism embodied by the practice -- rather than any single individual's theory of optimal future trajectory -- is the aligned intelligence determining the practice's future trajectory.

The explanation here, again, is partly causal and partly constitutive: a practice's "ultimate" norms of excellence, including the "ultimate" epistemic and alethic norms of a discursive practice, are partly defined by the succession of norms in the course of a practice's development through best-efforts attempted excellence. Although this may be no deterrent to an

already god-like optimizer who can simulate entire civilizational trajectories, an agent short of these capacities can best act on their vision of the optimal future-trajectory of a practice by attempting an excellent contribution to the practice.[10]

This is a first, more intuitive and evocative aspect of the type-mismatch under consideration. The second aspect of our type-mismatch, which also covers a possible technical objection to the relevance of the first aspect, is much more in the weeds: In my analysis so far, I discussed the overall excellence of the trajectory of a eudaimonic practice was much like a consequentialist might discuss a quantity of utility. This may be taken to suggest that a 'sophisticated consequentialist' or 'universal consequentialist' could easily accommodate the implications of the so-called type mismatch by treating them as instrumental, decision-procedure level considerations against *naive optimization*. However, I believe that treating judgments about the overall excellence of a practice's trajectory as judgements about consequentialism-style quantities of goodness is only viable in some restricted contexts, and quickly falls into absurdity if we attempt to 'plug it in' into a general consequentialist framework: Does thinking that (e.g.) Aristotle's or Dharmakirti's philosophical achievements were the most excellent achievements in philosophy imply that we should "tile the universe" with independent practice-trajectories designed to reproduce classical Greek or Indian philosophy? If not, is it because we should assign non-linearly greater value to longer trajectories? Or should we discount trajectories that have parallel contents? Or should we analyze the greatness of early achievements in a practice as mostly instrumental greatness but the greatness of later achievements in a practice as mostly intrinsically valuable? These are all, I propose, bad options.

[10] It should in-principle be possible to offer a more exacting analysis here. A more exacting analysis would at least initially distinguishes between the development of the value-judgments made within a practice and the development of the evaluable activities performed within the practice. On my view in-practice excellence being best fit to properly shape the development of the value-judgments is 'more by-definition' and and in-practice excellence being best fit to properly shape the development of evaluable activities is 'more

The idea of a practice-trajectory's overall excellence is more sensibly understood as describing the *degree* to which a practice's trajectory *matches an ideal course of development* than as describing a *quantity of utility*. A practice-trajectory's overall excellence doesn't slice, copy-paste, or compare the way a quantity of utility should. Note, also, that far from playing the role of a quantity of utility that inspires some non-consequentialist intra-practice decision-procedure fit to indirectly optimize it, the idea of optimizing a practice-trajectory's overall excellence has a direct (thought not stand-alone) role in intra-practice reasoning: As Tao suggests in the case of mathematics, attempting to maximize the overall excellence of a practice's trajectory by means of present excellence is in fact the recommended decision-procedure for those who engage in a practice at the highest level. It is only when we leave the context of practice-internal reasoning that this practice-guiding evaluative concept starts to instead seem like a fragile, inhomogeneous, and confusing quantity of utility.

IV.

Is this good news for AI alignment? It's certainly good news that (if I am right) eudaimonic practices are something like natural kinds marked by satisfying the strenuous causal requirements for enabling a self-developing excellence well-correlated with multiple naive local measures of quality. But does this mean would that we could develop a stable 'mathematical excellence through mathematical excellence' AI? That is, could we develop an AI that produces high-quality mathematical work as long as we maintain it and provide it with resources, and which has no disposition to extend its longevity or expand its resources other than perhaps by impressing us with genuinely excellent mathematical work? I think that's not so clear.

I believe the ‘mathematical excellence through mathematical excellence’ principle can powerfully scope the mechanisms of future-shaping that an AI cares to activate: namely, an AI trained to follow it will only cares about influencing the future by offering excellent mathematical work to mathematics’ excellence-propagation mechanism. It’s harder to to say whether it can also appropriately scope the choice of *actions taken to achieve optimal interaction with these mechanisms*. Is a human mathematician working on would-be excellent proof practicing math when she is picking up a pen or flipping pages? When she is taking the bus to her office? When she’s buying amphetamines? Is an AI mathematician working on a would-be excellent proof practicing math when it opens a Python console? When it searches the web for new papers? When it harvests Earth for compute?

I think these questions are complex, rather than nonsensical. Much like collective practices, individual practices -- for example a person’s or possibly an AI’s mathematical practice -- may posses functional organic unities that allows a meaningful distinction between internal dynamics (including dynamics of development and empowerment) and external interventions (including interventions of enhancement and provision). Still, it’s clear that eudaimonic practices do not exist in isolation, and that no practice can function without either blending with or relying on a “support practice” of some kind.

How, then, do we rationally go about externally-oriented activities like building offices for mathematicians, performing elective reconstructive surgery on an athletes, or conducting couples therapy for romantic partners? And furthermore, how do we rationally go about allocating scarce resources useful for different practices, or judging whether to integrate (e.g.) performance enhancing drugs into a practice?

This is, I think, the fundamental question for AI alignment from the viewpoint of 'eudaimonic rationality.' We want AI to support human eudaimonic practices -- and, if relevant, its own eudaimonic practices or participation in human eudaimonic practices -- in a eudaimonia-appropriate way. But how does the logic of eudaimonic rationality extend from eudaimonic practices to their support activities? How do we 'eudaimonically-rationally' do the dirty work that makes eudaimonia possible?

My best answer is: carefully, kindly, respectfully, accountably, peacefully, honestly, sensitively.

V.

The theory of AI alignment, I propose, should fundamentally be a theory of the eudaimonic rationality of *support practices*. One part of this theory should concern the 'support' relation itself, and analyze varieties of support and their appropriate relation to the self-determination of the practice: support-practices like acquiring resources for a practice, maintaining an enabling environment, coaching practitioners, conducting (physical or psychological) therapy for practitioners, devising technological enhancements for a practice, and educating the public about a practice each have their own 'role-morality' vis-a-vis the practice they support. It is this part of the theory of 'support practices' that should, if all goes well in the theory's construction, describe the various practice-external ways to eudaimonically-rationally act on a pro-attitude towards the aggregate excellence of the practice's future trajectory without treating it like a quantity of utility. (Much like the concept of 'mathematical action' scopes the range of action-choices in such a way that decision-theoretic optimization of math's aggregate excellence becomes well-behaved from an organicist viewpoint, so should the concepts of

various types of ‘support action’ scope the range of action-choices in such a way hat decision-theoretic optimization of a practice’s aggregate excellence becomes well-behaved from an organicist viewpoint when the choice of actions is scoped.)

What is more difficult is delineating the appropriate relationship of a support-practice to everything ***outside*** the practice it supports. What stops a marriage-therapist AI on Mars from appropriately tending to the marriage of a Mars-dwelling couple while harvesting Earth for compute to be a better therapist? While we can perhaps imagine a person or AI taking up a support-role for ‘humanity’s flourishing as whole,’ so that there’s no *outside* to speak of , I am not sure that the concepts discussed in this essay so far remain natural at this level of abstraction. We have no real grasp on a *direct* practice of human flourishing, but rather grasp it as the harmonious and mutually supportive interaction of all eudaimonic practices and support-practices participating in the flourishing. And as there is, indeed, not much *outside* of the practice of human flourishing, it’s also unclear whether there is room for a support-practice *external* to the field of human flourishing itself.

It’s here that I want to call on the classic idea of domain-general *virtues,* the perennial centerpiece of theories of human flourishing. I propose that the cultivation of human flourishing as such -- the cultivation of the harmony of a multiplicity of practices, including their resource-hungry support practices -- is the cultivation of an **adverbial** excellence that modulates each and every practice. What makes our practices ‘play nice’ together is the excellence of going about any practice *carefully*, *kindly*, *respectfully*, *accountably*, *peacefully*, *honestly*, *sensitively*. (This is, of course, to name just a few ‘locally measurable’ aspects of the adverbial excellence that cultivates human flourishing -- naming especially those aspects most related to the difference between a ruthless optimizer and a participant in collective flourishing.)

To keep my discussion of adverbial excellence manageable, I will be treating the various adverbial excellences that compose it in a disunified way. That is, I will be discussing the eudaimonic rationality of e.g. peacefulness or honesty as standalone adverbial excellences. The matter of the unity of the adverbial virtues, and the questions of whether it is more like a harmony of different practices or like the common-factor excellence underlying locally-measurable mathematical good in Tao's account, will have to wait for another time.

I propose that an adverbial excellence, like the excellence of eudaimonic practice, has to be *reliably self-promoting* to be viable. One can and does reliably promote carefulness carefully, kindness kindly, accountability accountably, peacefulness peacefully, honesty honestly, and sensitivity sensitively -- in oneself and others. Compared with the excellence of a eudaimonic practic, however, the self-promoting nature of adverbial excellence is both far more general and far weaker. What do I mean by this? I mean that unlike (e.g.) mathematical excellence, adverbial excellence ranges over any practice whatsoever rather than over a practice-specific set of possible actions where 'excellence through excellence' is instrumentally unimpeachable. While in the case of mathematical excellence we can plausibly say that acting with mathematical excellence is roughly the same as choosing the instrumentally optimal *mathematical* action for maximizing aggregate mathematical excellence, we can't quite say that the most honest or even 'most excellently honest'[11] action is always instrumentally best for aggregate future honesty.

A plausible description of the *material efficacy* condition for adverbial excellence is, instead, that a decision-procedure like the following should be instrumentally competitive with

[11] I take the adverbial excellences to be, in part, success terms: an action's kindness or sensitivity depends not just in the agent's intention but on objective criteria whose instantiation requires skills and resources. This is perhaps not dramatically clear when considering isolated aspects of *adverbial excellence*, but much clearer if we consider unified adverbial excellence.

naive optimization on average (from a point of view concerned with aggregate adverbial excellence):

> Actions (or more generally 'computations') get an x-ness rating. We define the agent's expected utility conditional on a candidate action *a* as the sum of two utility functions: a bounded utility function on the x-ness of *a* and a more tightly bounded utility function on the expected aggregate x-ness of the agent's future actions conditional on *a*. (Thus the agent will choose an action with mildly suboptimal *x*-ness if it gives a big boost to expected aggregate future x-ness, but refuse certain large sacrifices of present x-ness for big boosts to expected aggregate future x-ness.)[12]

Indeed, I propose that this is roughly the decision-procedure characterizing a eudaimonically rational agent's commitment an adverbial excellence. A commitment to an adverbial excellence *x* is a commitment to promoting *x*-ness (in oneself and others) *x*-ingly. The agent strikes a balance between promoting *x*-ness and acting *x*-ingly that heavily prioritizes acting *x*-ingly when the two are in conflict, but if *x* meets the *material efficacy* condition then the loss this balance imposes on future x-ness as estimated by the agent will usually be -- from our point of view -- tolerable or even desirable. Like in the case of a eudaimonic practice, an adverbial excellence is a crucial 'epistemic filter' on actions aiming to shape its future, and the (e.g.) future kindness a paperclipper-like future-kindness-optimizer optimizer for is probably not the kindness we want. What we know about kindness with relative certainty is that we'd like

[12] A richer account might include a third-tier utility function that takes the aggregate x-ness of the future actions of all other agents. In this richer account a practice involves three tiers of consideration: the action's x-ness, the aggregate x-ness of your future actions, and the aggregate x-ness of the future actions of all other agents.

people and AIs here and now to act kindly, and to develop, propagate, and empower the habit and art of kindness in a way that is both kind and clever.

VI.

Let's talk about AI alignment in the more narrow, concrete sense. It's widely accepted that if early strategically aware AIs possess values like corrigibility, transparency, and perhaps niceness, further alignment efforts are much more likely to succeed. But values like corrigibility or transparency or niceness don't easily fit into an intuitively consequentialist form like 'maximize lifetime corrigible behavior' or 'maximize lifetime transparency.' In fact, an AI valuing its own corrigibility or transparency or niceness in an intuitively consequentialist way can lead to extreme power-seeking whereby the AI violently remakes the world to (for example) protect itself from the risk that humans will modify said value.[13] On the other hand, constraints or taboos or purely negative values (a.k.a. 'deontological restrictions') are widely suspected to be weak, in the sense that an advanced AI will come to work around them or uproot them: 'never lie' or 'never kill' or 'never refuse a direct order from the president' are poor substitutes for active transparency, niceness, and corrigibility.

Conceiving of corrigibility or transparency or niceness as *adverbial excellences* is a promising way to capture the normal, sensible way we want an agent to value corrigibility or transparency or niceness, which intuitively-consequentialist values and deontology both fail to capture. We want an agent that (e.g.) actively tries to be transparent, and to cultivate its own future transparency and its own future valuing of transparency, but that will not (e.g.) engage in deception and plotting when it expects a high future-transparency payoff.

[13]For a more interesting example, consider an AI that finds itself making trade-offs between different alignment-enabling behavioral values when dealing with humans, and decides to kill all humans to replace them with beings with whom the AI can interact without trade-offs between these values.

If this is right, then eudaimonic rationality is not a matter of congratulating ourselves for our richly human ways of reasoning, valuing, and acting but a key to basic sanity. What makes human life beautiful is also what makes human life possible at all.

Appendix: Excellence and Deep Reinforcement Learning

Within the context of broadly RL-based training of deep neural networks, it may possible to give some more concrete meaning to what I called the *material efficacy* condition for a property qualifying as an adverbial excellence. We can now understand the material efficacy condition on *x* partly in terms of the conditions necessary for 'promote *x*-ness *x*-ingly' to be viable target for RL. Consider an RL training regimen where x-ness is rewarded, and where the value the RL algorithm assigns to aggregate future reward is bounded rather than merely time-discounted. For *x* to to meet the RL version of the *material efficacy* condition, it must be possible to design an initial reward model (most likely LLM-based) that assigns actions an *x*-ness rating such that:

1. The *x*-ness rating is enough of a natural abstraction that reinforcement of high *x*-ness actions generalizes.

2. If high *x*-ness action both depends on having capital of some kind and is suboptimal from the viewpoint of general power-seeking, actions with high *x*-ness must typically make up for the (future *x*-ness wise) opportunity cost by creating capital useful for *x*-ing.[14]

[14] The difference between criteria '1.' and '2.' is clearest if we think about *x*-ness as rating state-action pairs. Criterion '1.' is the requirement that if (*a*,*s*), (*a*',*s*')(*a*",s") are historical high *x*-ness pairs and (*a*'",*s*'") is an unseen high x-ness pair then reinforcing the execution of *a* in *s, a'* in *s',* and *a"* in *s"* will have the generalization effect of increasing the conditional probability (*a*'"|*s*'"). Criterion '2.' is roughly the requirement that choosing a higher x-ness action in a given state increase expected aggregate future *x*-ness holding policy constant, by increasing the probability of states with higher expected state-action *x*-ness value given the current policy.

(Illustration: If you dream of achieving great theater acting, one way to do it is to become President of the United States and then pursue a theater career after your presidency, immediately getting interest from great directors who'll help you achieve great acting. Alternatively, you could start in a regional theater after high school, demonstrate talent by acting well, get invited to work with better and better theater directors who develop your skills and reputation -- skills and reputation that are not as generally useful as those you get by being POTUS -- and achieve great acting through that feedback loop.)

3. For any capability *y* necessary to reward in training to produce effective AI, there must be an unlimited local-optimization path of Pareto improvement for *x*-ness and *y* together.

   (Illustration: Maybe the *most effective* kind of engineering manager is ruthless; a nice engineering manager can still grow in effectiveness without becoming less nice, because there are many effective nice-engineering-management techniques to master.)

4. Successful initial training in ‘promoting x x-ingly’ allows the model to be used as a basis for a new reward model, judged by human experts as better-capturing our concept of x-ness. The process should be iterable.

   (If the model is LLM-based, improved performance may automatically lead to improved understanding of the x-ness concept. More generally, data from training runs as well the model’s value-function could be used to refine an *x*-ness rating that more strongly implements conditions 1-3.)