The Universal Plot: Part I – Consciousness vs. Pure Replicators

Andrés Gómez Emilsson

December 2017

"It seems plain and self-evident, yet it needs to be said: the isolated knowledge obtained by a group of specialists in a narrow field has in itself no value whatsoever, but only in its synthesis with all the rest of knowledge and only inasmuch as it really contributes in this synthesis toward answering the demand, 'Who are we?'"

– Erwin Schrödinger in *Science and Humanism* (1951)

"Should you or not commit suicide? This is a good question. Why go on? And you only go on if the *game is worth the candle*. Now, the universe has been going on for an incredibly long time. Really, a satisfying theory of the universe should be one that's worth betting on. That seems to me to be absolutely elementary common sense. If you make a theory of the universe which isn't worth betting on... why bother? Just commit suicide. But if you want to go on playing the game, you've got to have an optimal theory for playing the game. Otherwise there's no point in it."

–Alan Watts, talking about Camus' claim that suicide is the most important question (cf. The Most Important Philosophical Question)

In this article we provide a novel framework for ethics which focuses on the perennial battle between **wellbeing-oriented consciousness-centric values** and **valueless patterns who happen to be great at making copies of themselves** (aka. *Consciousness vs. Pure Replicators*). This framework extends and generalizes modern accounts of ethics and intuitive wisdom, making intelligible numerous paradigms that previously lived in entirely different worlds (e.g. incongruous aesthetics and cultures). We place this worldview within a novel *scale of ethical development* with the following levels: (a) The Battle Between Good and Evil, (b) The Balance Between Good and Evil, (c) Gradients of Wisdom, and finally, the view that we advocate: (d) Consciousness vs. Pure Replicators. More so, we analyze each of these worldviews in light of our philosophical background assumptions and posit that (a), (b), and (c) are, at least in spirit, approximations to (d), except that they are less lucid, more confused, and

liable to exploitation by pure replicators. Finally, we provide a mathematical formalization of the problem at hand, and discuss the ways in which different theories of consciousness may affect our calculations. We conclude with a few ideas for how to avoid particularly negative scenarios.

1 Introduction

Throughout human history, the big picture account of the nature, purpose, and limits of reality has evolved dramatically. All religions, ideologies, scientific paradigms, and even aesthetics have background philosophical assumptions that inform their worldviews. One's answers to the questions "what exists?" and "what is good?" determine the way in which one evaluates the merit of beings, ideas, states of mind, algorithms, and abstract patterns.

Kuhn's claim that different scientific paradigms are mutually unintelligible (e.g. consciousness realism vs. reductive eliminativism) can be extended to worldviews in a more general sense. It is unlikely that we'll be able to convey the Consciousness vs. Pure Replicators paradigm by justifying each of the assumptions used to arrive to it one by one starting from current ways of thinking about reality. This is because these background assumptions support each other and are, individually, not derivable from current worldviews. They need to appear together as a unit to hang together tight. Hence, we now make the jump and show you, without further due, all of the background assumptions we need:

1. Consciousness Realism
2. Qualia Formalism
3. Valence Structuralism
4. The Pleasure Principle (and its corollary The Tyranny of the Intentional Object)
5. Physicalism (in the causal sense)
6. Open Individualism (also compatible with Empty Individualism)
7. Universal Darwinism

These assumptions have been discussed in previous articles. In the meantime, here is a brief description: (1) is the claim that consciousness is an element of reality rather than simply the improper reification of illusory phenomena, such that your conscious experience right now is as much a factual and determinate aspect of reality as, say, the rest mass of an electron. In turn, (2) qualia formalism is the notion that consciousness is in principle quantifiable. Assumption (3) states that valence (i.e. the pleasure/pain axis, how good an experience feels) depends of the structure of such experience (more formally, on the properties of the mathematical object isomorphic to its phenomenology).

(4) is the assumption that people's behavior is motivated by the pleasure-pain axis even when they think that's not the case. For instance, people may explicitly represent the reason for doing things in terms of concrete facts about the circumstance, and the pleasure principle does not deny that such reasons are important. Rather, it merely says that such reasons are motivating because one expects/anticipates less negative valence or more positive valence. The Tyranny of the Intentional Object describes the fact that we attribute changes in our valence to external events and objects, and believe that such events and objects are intrinsically good (e.g. we think "ice cream is great" rather than "I feel good when I eat ice cream").

Physicalism (5) in this context refers to the notion that the equations of physics fully describe the causal behavior of reality. In other words, the universe behaves according to physical laws and even consciousness has to abide by this fact.

Open Individualism (6) is the claim that we are all one consciousness, in some sense. Even though it sounds crazy at first, there are rigorous philosophical arguments in favor of this view. Whether this is true or not is, for the purpose of this article, less relevant than the fact that we can experience it as true, which happens to have both practical and ethical implications for how society might evolve.

Finally, (7) Universal Darwinism refers to the claim that natural selection works at every level of organization. The explanatory power of evolution and fitness landscapes generated by selection pressures is not confined to the realm of biology. Rather, it is applicable all the way from the quantum foam to, possibly, an ecosystem of universes.

The power of a given worldview is not only its capacity to explain our observations about the inanimate world and the quality of our experience, but also in its capacity to explain *in its own terms* the reasons for why other worldviews are popular as well. In what follows we will utilize these background assumptions to evaluate other worldviews.

2 The Four Worldviews About Ethics

The following four stages describe a plausible progression of thoughts about ethics and the question "what is valuable?" as one learns more about the universe and philosophy. Despite the similarity of the first three levels to the levels of other scales of moral development (e.g. this, this, this, etc.), we believe that the fourth level is novel, understudied, and very, very important.

2.1 The "Battle Between Good and Evil" Worldview

"Every distinction wants to become the distinction between good and evil." – Michael Vassar (source)

Common-sensical notions of *essential* good and evil are pre-scientific. For reasons too complicated to elaborate on for the time being, the human mind is capable of evoking an agentive sense of ultimate goodness (and of ultimate evil).

Figure 1: Good vs. Evil? God vs. the Devil?

Children are often taught that there are good people and bad people. That evil beings exist objectively, and that it is righteous to punish them and see them with scorn. On this level people reify anti-social behaviors as *sins*.

Essentializing good and evil, and tying it up to entities seems to be an early developmental stage of people's conception of ethics, and many people end up perpetually stuck in here. Several religions (specially the Abrahamic ones) are often practiced in such a way so as to reinforce this worldview. That said, many ideologies take advantage of the fact that a large part of the population is at this level to recruit adherents by redefining "what good and bad is" according to the needs of such ideologies. As a psychological attitude (rather than as a theory of the universe), reactionary and fanatical social movements often rely implicitly on this way of seeing the world, where there are bad people (Jews, traitors, infidels, over-eaters, etc.) who are seen as corrupting the soul of society and who deserve to have their fundamental badness exposed and exorcised with punishment in front of everyone else.

Implicitly, this view tends to gain psychological strength from the background assumptions of Closed Individualism (which allows you to imagine that people can be essentially bad). Likewise, this view tends to be naïve about the importance of valence in ethics. Good feelings are often interpreted as the result of being aligned with fundamental goodness, rather than as positive states of consciousness that happen to be triggered by a mix of innate and programmable things (including cultural identifications). More so, good feelings that don't come in response to the preconceived universal order are seen as demonic and aberrant.

From our point of view (the 7 background assumptions above) we interpret this particular worldview as something that we might be biologically predisposed to buy into. Believing in the battle between good and evil was probably

Figure 2: Traditional notions of God vs. the Devil can be interpreted as the personification of positive and negative valence

evolutionarily adaptive in our ancestral environment, and might reduce many frictional costs that arise from having a more subtle view of reality (e.g. "The cheaper people are to model, the larger the groups that can be modeled well enough to cooperate with them." – Michael Vassar). Thus, there are often pragmatic reasons to adopt this view, specially when the social environment does not have enough resources to sustain a more sophisticated worldview. Additionally, at an individual level, creating strong boundaries around what is or not permissible can be helpful when one has low levels of impulse control (though it may come at the cost of reduced creativity).

On this level, explicit wireheading (whether done right or not) is perceived as either sinful (defying God's punishment) or as a sort of treason (disengaging from the world). Whether one feels good or not should be left to the whims of the higher order. On the flip side, based on the pleasure principle it is possible to interpret the desire to be righteous as being motivated by high valence states, and reinforced by social approval, all the while the tyranny of the intentional object cloaks this dynamic.

It's worth noting that cultural conservativism, low levels of the psychological constructs of openness to experience and tolerance of ambiguity, and high levels of need for closure, all predict getting stuck in this worldview for one's entire life.

2.2 The "Balance Between Good and Evil" Worldview

TVTropes has a great summary of the sorts of narratives that express this particular worldview and I highly recommend reading that article to gain insight into the moral attitudes compatible with this view. For example, here are some reasons why Good cannot or should not win:

> Good winning includes: the universe becoming boring, society stagnating or collapsing from within in the absence of something to struggle against or giving people a chance to show real nobility and virtue by risking their lives to defend each other. Other times, it's enforced by depicting ultimate good as repressive (often Lawful Stupid), or by declaring concepts such as free will or ambition as evil. In other words "too much of a good thing".
>
> – "Balance Between Good and Evil" by TV Tropes

Now, the stated reasons why people might buy into this view are rarely their true reasons. Deep down, the Balance Between Good and Evil is adopted because: people want to differentiate themselves from those who believe in the Battle Between Good and Evil to signal intellectual sophistication, they experience learned helplessness after trying to defeat evil without success (often in the form of resilient personal failings or societal flaws), they find the view compelling at an intuitive emotional level (i.e. they have internalized the hedonic treadmill and project it onto the rest of reality).

In all of these cases, though, there is something somewhat paradoxical about holding this view. And that is that people report that coming to terms with the fact that not everything can be good is itself a cause of relief, self-acceptance, and happiness. In other words, holding this belief is often mood-enhancing. One can also confirm the fact that this view is emotionally load-bearing by observing the psychological reaction that such people have to, for example, bringing up the Hedonistic Imperative (which asserts that eliminating suffering without sacrificing anything of value is scientifically possible), indefinite life extension, or the prospect of super-intelligence. Rarely are people at this level intellectually curious about these ideas, and they come up with excuses to avoid looking at the evidence, however compelling it may be.

For example, some people are lucky enough to be born with a predisposition to being hyperthymic (which, contrary to preconceptions, does the opposite of making you a couch potato). People's hedonic set-point is at least partly genetically determined, and simply avoiding some variants of the SCN9A gene with preimplantation genetic diagnosis would greatly reduce the number of people who needlessly suffer from chronic pain.

But this is not seen with curious eyes by people who hold this or the previous worldview. Why? Partly this is because it would be painful to admit that both oneself and others are stuck in a local maxima of well-being and that examining alternatives might yield very positive outcomes (i.e. omission bias). But at its core, this willful ignorance can be explained as a consequence of the fact that people at this level get a lot of positive valence from interpreting present and past suffering in such a way that it becomes tied to their core identity. Pride in having overcome their past sufferings, and personal attachment to their current struggles and anxieties binds them to this worldview.

If it wasn't clear from the previous paragraph, this worldview often requires a special sort of chronic lack of self-insight. It ultimately relies on a psychological trick. One never sees people who hold this view voluntarily breaking their legs,

taking poison, or burning their assets to increase the goodness elsewhere as an act of altruism. Instead, one uses this worldview as a mood-booster, and in practice, it is also susceptible to the same sort of fanaticism as the first one (although somewhat less so). "There can be no light without the dark. And so it is with magic. Myself, I always try to live within the light." – Horace Slughorn.

Additionally, this view helps people rationalize the negative aspects of one's community and culture. For example, it not uncommon for people to say that buying factory farmed meat is acceptable on the grounds that "some things have to die/suffer for others to live/enjoy life." Balance Between Good and Evil is a close friend of status quo bias.

Hinduism, Daoism, and quite a few interpretations of Buddhism work best within this framework. Getting closer to God and ultimate reality is not done by abolishing evil, but by embracing the unity of all and fostering a healthy balance between health and sickness.

Figure 3: Taijitu

It's also worth noting that the balance between good and evil tends to be recursively applied, so that one is not able to "re-define our utility function from 'optimizing the good' to optimizing 'the balance of good and evil' with a hard-headed evidence-based consequentialist approach." Indeed, trying to do this is then perceived as yet another incarnation of good (or evil) which needs to also be balanced with its opposite (willful ignorance and fuzzy thinking). One comes to the conclusion that it is the fuzzy thinking itself that people at this level are after: to blur reality just enough to make it seem good, and to feel like

one is not responsible for the suffering in the world (specially by inaction and lack of thinking clearly about how one could help). "Reality is only a Rorschach ink-blot, you know" – Alan Watts. So this becomes a justification for thinking less than one really has to about the suffering in the world. Then again, it's hard to blame people for trying to keep the collective standards of rigor lax, given the high proportion of fanatics who adhere to the "battle between good and evil" worldview, and who will jump the gun to demonize anyone who is slacking off and not stressed out all the time, constantly worrying about the question "could I do more?"

(Note: if one is actually trying to improve the world as much as possible, being stressed out about it all the time is not the right policy).

2.3 The "Gradients of Wisdom" Worldview

David Chapman's HTML book *Meaningness* might describe both of the previous worldviews as variants of *eternalism*. In the context of his work, eternalism refers to the notion that there is an absolute order and meaning to existence. When applied to codes of conduct, this turns into "ethical eternalism", which he defines as: "the stance that there is a fixed ethical code according to which we should live. The eternal ordering principle is usually seen as the source of the code." Chapman eloquently argues that eternalism has many side effects, including: deliberate stupidity, attachment to abusive dynamics, constant disappointment and self-punishment, and so on. By realizing that, in some sense, no one knows what the hell is going on (and those who do are just pretending) one takes the first step towards the "Gradients of Wisdom" worldview.

At this level people realize that there is no evil essence. Some might talk about this in terms of there "not being good or bad people", but rather just degrees of impulse control, knowledge about the world, beliefs about reality, emotional stability, and so on. A villain's soul is not connected to some kind of evil reality. Rather, his or her actions can be explained by the causes and conditions that led to his or her psychological make-up.

Sam Harris' ideas as expressed in *The Moral Landscape* evoke this stage very clearly. Sam explains that just as health is a fuzzy but important concept, so is psychological wellbeing, and that for such a reason we can objectively assess cultures as more or less in agreement with human flourishing.

Indeed, many people who are at this level do believe in valence structuralism, where they recognize that there are states of consciousness that are inherently better in some intrinsic subjective value sense than others.

However, there is usually no principled framework to assess whether a certain future is indeed optimal or not. There is little hard-headed discussion of population ethics for fear of sounding unwise or insensitive. And when push comes to shove, they lack good arguments to decisively rule out why particular situations might be bad. In other words, there is room for improvement, and such improvement might eventually come from more rigor and bullet-bitting. In particular, a more direct examination of the implications of: Open Individualism, the Tyranny of the Intentional Object, and Universal Darwinism can allow

Figure 4: *The Moral Landscape*

someone on this level to make a breakthrough. Here is where we come to:

2.4 The "Consciousness vs. Pure Replicators" Worldview

In "Wireheading Done Right" we introduced the concept of a *pure replicator*:

> I will define a pure replicator, in the context of agents and minds, to be an intelligence that is indifferent towards the valence of its conscious states and those of others. A pure replicator invests all of its energy and resources into surviving and reproducing, even at the cost of continuous suffering to themselves or others. Its main evolutionary advantage is that it does not need to spend any resources making the world a better place.

Presumably our genes are pure replicators. But we, as sentient minds who recognize the intrinsic value (both positive and negative) of conscious experiences, are not pure replicators. Thanks to a myriad of fascinating dynamics, it so happened that making minds who love, appreciate, think creatively, and philosophize was a side effect of the process of refining the selfishness of our genes. We must not take for granted that we are more than pure replicators ourselves, and that we care both about our wellbeing and the wellbeing of others. The problem now is that the particular selection pressures that led to this may not be present in the future. After all, digital and genetic technologies are drastically changing the fitness landscape for patterns that are good at making copies of themselves.

In an optimistic scenario, future selection pressures will make us all naturally gravitate towards super-happiness. This is what David Pearce posits in his essay "The Biointelligence Explosion":

> As the reproductive revolution of "designer babies" gathers pace, prospective parents will pre-select alleles and allelic combinations

> for a new child in anticipation of their behavioural effects – a novel kind of selection pressure to replace the "blind" genetic roulette of natural selection. In time, routine embryo screening via preimplantation genetic diagnosis will be complemented by gene therapy, genetic enhancement and then true designer zygotes. In consequence, life on Earth will also become progressively happier as the hedonic treadmill is recalibrated. In the new reproductive era, hedonic set-points and intelligence alike will be ratcheted upwards in virtue of selection pressure. For what parent-to-be wants to give birth to a low-status depressive "loser"? Future parents can enjoy raising a normal transhuman supergenius who grows up to be faster than Usain Bolt, more beautiful than Marilyn Monroe, more saintly than Nelson Mandela, more creative than Shakespeare – and smarter than Einstein.

In a pessimistic scenario, the selection pressures lead to the opposite direction, where negative experiences are the only states of consciousness that happen to be evolutionarily adaptive, and so they become universally used.

There is a number of thinkers and groups who can be squarely placed on this level, and relative to the general population, they are extremely rare (see: The Future of Human Evolution, A Few Dystopic Future Scenarios, Book Review: Age of EM, Nick Land's Gnon, Spreading Happiness to the Stars Seems Little Harder than Just Spreading, etc.). See also.[1] What is much needed now, is formalizing the situation and working out what we could do about it. But first, some thoughts about the current state of affairs.

There is at least some encouraging facts that suggest it is not too late to prevent a pure replicator takeover. There are memes, states of consciousness, and resources that can be used in order to steer evolution in a positive directions. In particular, as of 2017:

1. A very big proportion of the economy is dedicated to trading positive experiences for money, rather than just survival or power tools. Thus an economy of information about states of consciousness is still feasible.

2. There is a large fraction of the population who is altruistic and would be willing to cooperate with the rest of the world to avoid catastrophic

[1] Related Work:

Here is a list of literature that points in the direction of Consciousness vs. Pure Replicators. There are countless more worthwhile references, but I think that these ones are about the best:

The Biointelligence Explosion (David Pearce), Meditations on Moloch (Scott Alexander), What is a Singleton? (Nick Bostrom), Coherent Extrapolated Volition (Eliezer Yudkowsky), Simulations of God (John Lilly), Meaningness (David Chapman), The Selfish Gene (Richard Dawkins), Darwin's Dangerous Idea (Daniel Dennett), Prometheus Rising (R. A. Wilson).

Additionally, here are some further references that address important aspects of this worldview, although they are not explicitly trying to arrive at a big picture view of the whole thing:

Neurons Gone Wild (Kevin Simler), The Age of EM (Robin Hanson), The Mating Mind (Geoffrey Miller), The Joyous Cosmology (Alan Watts), The Ego Tunnel (Thomas Metzinger), The Orthogonality Thesis (Stuart Armstrong).

scenarios.

3. Happy people are more motivated, productive, engaged, and ultimately, economically useful (see hyperthymic temperament).

4. Many people have explored Open Individualism and are interested (or at least curious) about the idea that we are all one.

5. A lot of people are fascinated by psychedelics and the non-ordinary states of consciousness that they induce.

6. MDMA-like consciousness is both very positive in terms of its valence, but also, amazingly, extremely pro-social, and future sustainable versions of it could be recruited to stabilize societies where the highest value is the collective wellbeing.

It is important to not underestimate the power of the facts laid out above. If we get our act together and create a Manhattan Project of Consciousness we might be able to find sustainable, reliable, and powerful methods that stabilize a hyper-motivated, smart, super-happy and super-prosocial state of consciousness in a large fraction of the population. In the future, we may all by default identify with consciousness itself rather than with our bodies (or our genes), and be intrinsically (and rationally) motivated to collaborate with everyone else to create as much happiness as possible as well as to eradicate suffering with technology. And if we are smart enough, we might also be able to solidify this state of affairs, or at least shield it against pure replicator takeovers.

The beginnings of that kind of society may already be underway. Consider for example the contrast between Burning Man and Las Vegas. Burning Man is a place that works as a playground for exploring post-Darwinean social dynamics, in which people help each other overcome addictions and affirm their commitment to helping all of humanity. Las Vegas, on the other hand, might be described as a place that is filled to the top with pure replicators in the forms of memes, addictions, and denial. The present world has the potential for both kind of environments, and we do not yet know which one will outlive the other in the long run.

3 Formalizing the Problem

We want to specify the problem in a way that will make it mathematically intelligible. In brief, in this section we focus on specifying what it means to be a pure replicator in formal terms. Per the definition, we know that pure replicators will use resources as efficiently as possible to make copies of themselves, and will not care about the negative consequences of their actions. And in the context of using brains, computers, and other systems whose states might have moral significance (i.e. they can suffer), they will simply care about the overall utility of such systems for whatever purpose they may require. Such utility will be a

function of both the accuracy with which the system performs it's task, as well as its overall efficiency in terms of resources like time, space, and energy.

Simply phrased, we want to be able to answer the question: Given a certain set of constraints such as energy, matter, and physical conditions (temperature, radiation, etc.), what is the amount of pleasure and pain involved in the most efficient implementation of a given predefined input-output mapping?

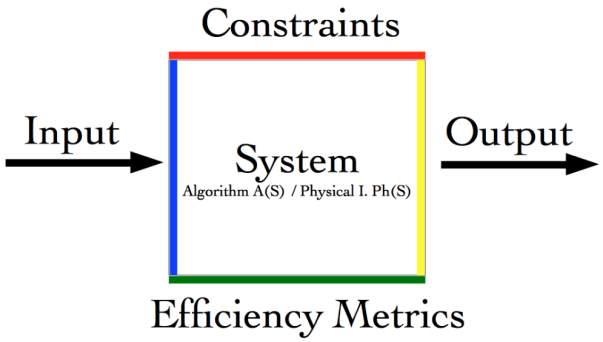

Figure 5: Input-output mapping

The image above represents the relevant components of a system that might be used for some purpose by an intelligence. We have the inputs, the outputs, the constraints (such as temperature, materials, etc.) and the efficiency metrics. Let's unpack this. In the general case, an intelligence will try to find a system with the appropriate trade-off between efficiency and accuracy. We can wrap up this as an "efficiency metric function", e(o—i, s, c) which encodes the following meaning: "e(o—i, s, c) = the efficiency with which a given output is generated given the input, the system being used, and the physical constraints in place."

$$s = system$$
$$c = constraints$$
$$i = input$$
$$o = output$$
$$e(o|i, s, c) = efficiency\ metric$$

Figure 6: Key

Now, we introduce the notion of the "valence for the system given a particular input" (i.e. the valence for the system's state in response to such an input). Let's call this v(s—i). It is worth pointing out that whether valence can be computed, and whether it is even a meaningfully objective property of a

system is highly controversial (e.g. "Measuring Happiness and Suffering"). Our particular take (at QRI) is that valence is a mathematical property that can be decoded from the mathematical object whose properties are isomorphic to a system's phenomenology (see: Principia Qualia: Part II – Valence, and also Quantifying Bliss. If so, then there is a matter of fact about just how good/bad an experience is. For the time being we will assume that valence is indeed quantifiable, given that we are working under the premise of valence structuralism (as stated in our list of assumptions). We thus define the overall utility for a given output as U(e(o—i, s, c), v(s—i)), where the valence of the system may or may not be taken into account. In turn, an intelligence is said to be altruistic if it cares about the valence of the system in addition to its efficiency, so that it's utility function penalizes negative valence (and rewards positive valence).

$$v(s|i) = \textit{valence of system } s \textit{ given input } i$$
$$U(e(o|i, s, c), v(s|i)) = \textit{the utility of efficiency } e, \textit{ and valence } v$$

Figure 7: Valence altruism

Now, the intelligence (altruistic or not) utilizing the system will also have to take into account the overall range of inputs the system will be used to process in order to determine how valuable the system is overall. For this reason, we define the expected value of the system as the utility of each input multiplied by its probability.

$$P(i) = \textit{probability of input } I$$
$$P(I) = \textit{probabilities for all inputs (abuse of notation)}$$

Figure 8: Input probabilities

(Note: a more complete formalization would also weight in the importance of each input-output transformation, in addition to their frequency). Moving on, we can now define the overall expected utility for the system given the distribution of inputs it's used for, its valence, its efficiency metrics, and its constraints as E[U(s—v, e, c, P(I))]:

$$E[U(s|v, e, c, P(I))] = \sum_{i \in I} P(i)U(e(o|i, s, c), v(s|i))$$
$$\arg\max_{s} E[U(s|v, e, c, P(I))] = \textit{the preferred system}$$

Figure 9: Chosen system

The last equation shows that the intelligence would choose the system that maximizes E[U(s—v, e, c, P(I))].

Pure replicators will be better at surviving as long as the chances of reproducing do not depend on their altruism. If altruism does not reduce such reproductive fitness, then:

Given two intelligences that are competing for existence and/or resources to make copies of themselves and fight against other intelligences, there is going to be a strong incentive to choose a system that maximizes the efficiency metrics regardless of the valence of the system.

In the long run, then, we'd expect to see only *non-altruistic intelligences* (i.e. intelligences with utility functions that are indifferent to the valence of the systems it uses to process information). In other words, as evolution pushes intelligences to optimize the efficiency metrics of the systems they employ, it also pushes them to stop caring about the wellbeing of such systems. In other words, evolution pushes intelligences to become pure replicators in the long run.

Hence we should ask: How can altruism increase the chances of reproduction? A possibility would be for the environment to reward entities that are altruistic. Unfortunately, in the long run we might see that environments that reward altruistic entities produce less efficient entities than environments that don't. If there are two very similar environments, one which rewards altruism and one which doesn't, the efficiency of the entities in the latter might become so much higher than in the former that they become able to takeover and destroy whatever mechanism is implementing such reward for altruism in the former. Thus, we suggest to find environments in which rewarding altruism is baked into their very nature, such that similar environments without such reward either don't exist or are too unstable to exist for the amount of time it takes to evolve non-altruistic entities. This and other similar approaches will be explored further in Part II.

4 Behaviorism, Functionalism, Non-Materialist Physicalism

A key insight is that the formalization presented above is agnostic about one's theory of consciousness. We are simply assuming that it's possible to compute the valence of the system in terms of its state. How one goes about computing such valence, though, will depend on how one maps physical systems to experiences. Getting into the weeds of the countless theories of consciousness out there would not be very productive at this stage, but there is still value in defining the rough outline of kinds of theories of consciousness. In particular, we categorize (physicalist) theories of consciousness in terms of the level of abstraction they identify as the place in which to look for consciousness.

Behaviorism and similar accounts simply associate consciousness to input-output mappings, which can be described, in Marr's terms, as the computational level of abstraction. In this case, v(s—i) would not depend on the details of the system as much as in what it does from a third person point of view. Behaviorists don't care what's in the Chinese Room; all they care about is if the Chinese

Room can scribble "I'm in pain" as an output. How we can formalize a mathematical equation to infer whether a system is suffering from a behaviorist point of view is beyond me, but maybe someone might want to give it a shot. As a side note, behaviorists historically were not very concerned about pain or pleasure, and there cause to believe that behaviorism itself might be antidepressant for people for whom introspection results in more pain than pleasure.

Functionalism (along with computational theories of mind) defines consciousness as the sum-total of the functional properties of systems. In turn, this means that consciousness arises at the algorithmic level of abstraction. Contrary to common misconception, functionalists do care about how the Chinese Room is implemented: contra behaviorists, they do not usually agree that a Chinese Room implemented with a look-up table is conscious.[2]

As such v(s—i) will depend on the algorithms that the system is implementing. Thus, as an intermediary step, one would need a function that takes the system as an input and returns the algorithms that the system is implementing as an output, A(s). Only once we have A(s) we would then be able to infer the valence of the system. Which algorithms, and for what reason, are in fact hedonically-charged has yet to be clarified. Committed functionalists often associate reinforcement learning with pleasure and pain, and one could imagine that as philosophy of mind gets more rigorous and takes into account more advancements in neuroscience and AI, we will see more hypothesis being made about what kinds of algorithms result in phenomenal pain (and pleasure). There are many (still fuzzy) problems to be solved for this account to work even in principle. Indeed, there is a reason to believe that the question "what algorithms is this system performing?" has no definite answer, and it surely isn't frame-invariant in the same way that a physical state might be. The fact that algorithms do not carve nature at its joints would imply that consciousness is not really a well-defined element of reality either. But rather than this working as a reductio-ad-absurdum of functionalism, many of its proponents have instead turned around to conclude that consciousness itself is not a natural kind. This does represent an important challenge in order to define the valence of the system, and makes the problem of detecting and avoiding pure replicators extra challenging. Admirably so, this is not stopping some from trying anyway.

We also should note that there are further problems with functionalism in general, including the fact that qualia, the binding problem, and the causal role of consciousness seem underivable from its premises. For a detailed discussion about this, read this article.

Finally, Non-Materialist Physicalism locates consciousness at the *implementation level of abstraction*. This general account of consciousness refers to the notion that the intrinsic nature of the physical is qualia. There are many related

[2]Rather, they usually claim that, given that a Chinese Room is implemented with physical material from this universe and subject to the typical constraints of this world, it is extremely unlikely that a universe-sized look-up table would be producing the output. Hence, the algorithms that are producing the output are probably highly complex and using information processing with human-like linguistic representations, which means that, by all means, the Chinese Room is very likely understanding what it is outputting.

views that for the purpose of this article should be good enough approximations: panpsychism, panexperientialism, neutral monism, Russellian monism, etc. Basically, this view takes seriously both the equations of physics and the idea that what they describe is the behavior of qualia. A big advantage of this view is that there is a matter-of-fact about what a system is composed of. Indeed, both in relativity and quantum mechanics, the underlying nature of a system is *frame-invariant*, such that its fundamental (intrinsic and causal) properties do not depend on one's frame of reference. In order to obtain v(s—i) we will need to obtain this frame-invariant description of what the *system is* in a given state. Thus, we need a function that takes as input physical measurements of the system and returns the best possible approximation to what is actually going on under the hood, Ph(s). And only with this function Ph(s) we would be ready to compute the valence of the system. Now, in practice we might not need a plank-length description of the system, since the mathematical property that describes it's valence might turn out to be well-approximated with high-level features of it.

The main problem with Non-Materialist Physicalism comes when one considers systems that have similar efficiency metrics, are performing the same algorithms, and look the same in all of the relevant respects from a third-person point, and yet do not have the same experience. In brief: if physical rather than functional aspects of systems map to conscious experiences, it seems likely that we could find two systems that *do the same* (input-output mapping), *do it in the same way* (algorithms), and yet one is conscious and the other isn't.

This kind of scenario is what has pushed many to conclude that functionalism is the only viable alternative, since at this point consciousness would seem epiphenomenal (e.g. Zombies Redacted). And indeed, if this was the case, it would seem to be a mere matter of chance that our brains are implemented with the right stuff to be conscious, since the nature of such stuff is not essential to the algorithms that actually end up processing the information. You cannot speak to stuff, but you can speak to an algorithm. So how do we even know we have the right stuff to be conscious?

The way to respond to this very valid criticism is for Non-Materialist Physicalism to postulate that bound states of consciousness have computational properties. In brief, epiphenomenalism cannot be true. But this does not rule out Non-Materialist Physicalism for the simple reason that the quality of states of consciousness might be involved in processing information. Enter...

5 The Computational Properties of Consciousness

Let's leave behaviorism behind for the time being. In what ways do functionalism and non-materialist physicalism differ in the context of information processing? In the former, consciousness is nothing other than certain kinds of information processing, whereas in the latter conscious states can be used for

information processing. An example of this falls out of David Pearce's theory of consciousness seriously. In his account, the phenomenal binding problem (i.e. "if we are made of atoms, how come our experience contains many pieces of information at once?", see: The Combination Problem for Panpsychism) is solved via quantum coherence. Thus, a given moment of consciousness is a definite physical system that works as a unit. Conscious states are *ontologically unitary*, and not merely *functionally unitary*.

If this is the case, there would be a good reason for evolution to recruit conscious states to process information. Simply put, given a set of constraints, using quantum coherence might be the most efficient way to solve some computational problems. Thus, evolution might have stumbled upon a computational jackpot by creating neurons whose (extremely) fleeting quantum coherence could be used to solve constraint satisfaction problems in ways that would be more energetically expensive to do otherwise. In turn, over many millions of years, brains got really good at using consciousness in order to efficiently process information. It is thus not an accident that we are conscious, that our conscious experiences are unitary, that our world-simulations use a wide range of qualia varieties, and so on. All of these seemingly random, seemingly epiphenomenal, aspects of our existence happen to be computationally advantageous. Just as using quantum computing for factorizing prime numbers, or for solving problems amenable to annealing might give quantum computers a computational edge over their non-quantum counterparts, so is using bound conscious experiences helpful to outcompete non-sentient animals.

Of course, there is yet no evidence of macroscopic decoherence and the brain is too hot anyway anyway, so on the face of it Pearce's theory seems exceedingly unlikely. But its explanatory power should not be dismissed out of hand, and the fact that it makes empirically testable predictions is noteworthy (how often do consciousness theorists make precise predictions to falsify their theories?).

Whether it is via quantum coherence, entanglement, invariants of the gauge field, or any other deep physical property of reality, non-materialist physicalism can avert the spectre of epiphenomenalism by postulating that the relevant properties of matter that make us conscious are precisely those that give our brains a computational edge (relative to what evolution was able to find in the vicinity of the fitness landscape explored in our history).

6 Will Pure Replicators Use Valence Gradients at All?

Whether we work under the assumption of functionalism or non-materialist physicalism, we already know that our genes found happiness and suffering to be evolutionary advantageous. So we know that there is at least a set of constraints, efficiency metrics, and input-output mappings that make both phenomenal pleasure and pain very good algorithms (functionalism) or physical implementations (non-materialist physicalism). But will the parameters neces-

sitated by replicators in the long-term future have these properties? Remember that evolution was only able to explore a restricted state-space of possible brain implementations delimited by the pre-existing gene pool (and the behavioral requirements provided by the environment). So, in one extreme case, it may be the case that a fully optimized brain simply does not need consciousness to solve problems. And in another extreme, it may turn out that consciousness is extraordinarily more powerful when used in an optimal way. Would this be good or bad?

What's the best case scenario? Well, the absolute best possible case is a case so optimistic and incredibly lucky that if it turned out to be true, it would probably make me believe in a benevolent God (or Simulation). This is the case where it turns out that only *positive valence gradients* are computationally superior to every other alternative given a set of constraints, input-output mappings, and arbitrary efficiency functions. In this case, the most powerful pure replicators, despite their lack of altruism, will nonetheless be pumping out massive amounts of systems that produce unspeakable levels of bliss. It's as if the very nature of this universe is blissful... we simply happen to suffer because we are stuck in a tiny wrinkle at the foothills of the optimization process of evolution.

In the extreme opposite case, it turns out that only *negative valence gradients* offer strict computational benefits under heavy optimization. This would be Hell. Or at least, it would tend towards Hell in the long run. If this happens to be the universe we live in, let's all agree to either conspire to prevent evolution from moving on, or figure out the way to *turn it off*. In the long term, we'd expect every being alive (or AI, upload, etc.) to be a zombie or a piece of dolorium. Not a fun idea.

In practice, it's much more likely that both positive and negative valence gradients will be of some use in some contexts. Figuring out exactly which contexts these are might be both extremely important, and also extremely dangerous. In particular, finding out in advance which computational tasks make positive valence gradients a superior alternative to other methods of doing the relevant computations would inform us about the sorts of cultures, societies, religions, and technologies that we should be promoting in order to give this a push in the right direction (and hopefully out-run the environments that would make negative valence gradients adaptive).

Unless we create a Singleton early on, it's likely that by default all future entities in the long-term future will be non-altruistic pure replicators. But it is also possible that there are multiple attractors (i.e. evolutionarily stable ecosystems) in which different computational properties of consciousness are adaptive. Thus the case for pushing our evolutionary history in the right direction right now before we give up.

7 Coming Next: The Hierarchy of Cooperators

Now that we covered the four worldviews, formalized what it means to be a pure replicator, and analyzed the possible future outcomes based on the computational properties of consciousness (and of valence gradients in particular), we are ready to face the game of reality in its own terms.

Team Consciousness, we need to to get our act together. We need a systematic worldview, availability of states of consciousness, set of beliefs and practices to help us prevent pure replicator takeovers.

But we cannot do this as long as we are in the dark about the sorts of entities, both consciousness-focused and pure replicators, who are likely to arise in the future in response to the selection pressures that cultural and technological change are likely to produce. In Part II of "The Universal Plot" we will address this and more. Stay tuned...