Title
Altruism: Past, Present, Propagation

Author
Caleiro, Diego

Publication Date
2021

Peer reviewed|Thesis/dissertation

Altruism: Past, Present, Propagation

By

Diego Caleiro

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Anthropology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge

Professor Terrence Deacon, Chair
Professor Alexei Yurchak
Professor Lawrence Cohen
Professor Dacher Keltner

Fall 2021

Abstract

Altruism: Past, Present, Propagation

by

Diego Caleiro

Doctor of Philosophy in Anthropology

University of California, Berkeley

Professor Terrence Deacon, Chair

I intend to examine the question of scalability of altruism and its long term sustainability. Altruism has posed multiple substantial conundrums in different fields, including challenges to its possibility, its efficiency, and its scalability. Multiple accounts of the evolution and dynamics involved in altruistic interaction among humans, institutions, animals and artificial intelligences rely on conflict between entities as the proximal cause of cooperation and altruism within entities. In this examination I go through our knowledge in anthropology, philosophy, neuroscience, evolutionary psychology, and biology to examine in which ways specific properties of humans and human groups, as well as theorized properties of agents in general, can accommodate truly altruistic actions and behaviors. Further I examine the question of where the processes that lead us to the current levels and type of altruism are headed and whether there is a basin of attraction towards which it would be desirable to go from an altruistic perspective in longer timescales. I will examine the limitations and constraints posed by game theory, neuroscience, and the specifics of our evolutionary past, and in so doing I paint a picture of altruism as a valuable and feasible, as well as scalable, strategy, for the foreseeable future. I lastly propose some roads toward achieving these scalable possibilities through a combination of evolutionary and technological nudges according to the current literature.

Table of Contents

Acknowledgements

I'm sure if I tried to list the over a hundred people who were pivotal and essential to the creation of this dissertation, from effective altruists to the board of advisers, from colleagues in research to old mentors, from family members to intellectual influences, I would forget some of pivotal importance and blame myself for decades for having forgotten. So I shall instead be grateful to all those who were causally influential in the creation of this work, whether or not they were stored in my memory as having done so. Thank you!

Introduction

Could Altruism Win?

A world in which global coordination problems remain unsolved even as the power of technology increases towards its physical limits is a world that is hostage to the possibility that – at some level of technological development – nature too strongly favours destruction over creation. - Nick Bostrom (2017)

Is Altruism a Scalable and Stable Strategy at Different Levels of Selection?

The question above is the object of scrutiny of the present work. Through this work I wish to examine both the birds eye view of altruism and cooperation at different scales, as well as the first person perspective of how to guide one's actions in altruistic fashion according to different areas such as anthropology, evolutionary psychology, philosophy, social psychology and others. The examination will include several subquestions:

- What is altruism according to different fields?
- Are humans altruistic and in what sense?
- Are some subgroups of humans more altruistic than others and what caused these differences?
- What are some possible stories for the emergence of altruism in our species?
- What are the cognitive mechanisms that we utilize to perform altruistic actions, that is, what are the gears operating behind altruistic acts?
- Are high levels of altruism stable in biology?
- Is altruism scalable and if so to what point?

- If we extrapolate currently known dynamics in biocultural evolution, should we anticipate altruism to continue stable, subside, or increase?
- What are the preconditions for the Moral cognition we utilize when deciding to act altruistically?
- How do different psychological modalities (System 1 and System 2) perform when attempting altruistic actions?
- In what way is Effective Altruism a movement that exemplifies general altruism, and in which ways it is an improvement or hindrance to previous evolved mechanisms of altruism?
- What should a deliberate altruist do under different scenarios so as to maximize impact of action?
- What should a deliberate group of altruists do under different scenarios to maximize the positive altruistic outcomes of their actions, and how does that compare with altruistic notions pertaining to behavioral genomics, economics, game theory, neuroscience, etc...
- By the coalition of natural, evolved, nudged, and deliberately steered forces, to what extent can altruism be expanded and scaled into the far future?

These questions can be considered sub-questions to the larger question of the stability or not of altruism in the long run. They are either literal subcomponents of the larger question, or they are premises that need to be settled in one or another direction to make an answer to the larger scalability question even answerable in principle.

We will examine the definition of altruism in different fields, including extensive examination of two concepts that cannot be left outside the scope of any investigation of altruism: Groups and Individuals.

The main reason these two concepts cannot be ignored is that throughout evolution there have been transitions from groups to individuals and vice-versa and these are preceded by periods of intense altruistic activity between members of a group that would eventually become individual like. A related set of notions: Organisms and Superorganisms will also be investigated, in particular with regards to human groups. A superorganism by definition contains altruistic and synergistic constituent elements, so wherever biocultural evolution formed superorganisms, it has in doing so formed somewhat stable altruistic clusters.

There are external and internal constraints to the emergence and permanence of altruism over time in individuals or groups:

> The external constraints can be mathematically expressed and are related to population dynamics, multilevel selection, and other evolutionary mechanisms that pose hard or soft barriers to the extent to which an organism can be altruistic under different circumstances.

> The internal ones are part of what Dennett would call the physical and intentional stances, they are related to the specifics of how our (or an animal's) moral cognition operates, which psychological mechanisms are utilized when undertaking altruistic actions.

There are also complexity and computational limits to altruistic actions in practice for any finite system attempting to act in other benefiting ways. I will not take much of this examination to discuss these as they are not fundamentally distinct from other discussions about finite computation, the scope of utilitarianism, turing machines and halting algorithms, which belong more closely to research in the philosophy of computer science and mathematics.

I will briefly touch on the topic of artificial agent altruism (as opposed to biological and biocultural agents) but mostly to cast light on how we as biocultural agents can act in more or less altruistic ways as well as to have an abstract model of altruism in systems with deterministic consequences.

I will also create a classification scheme for multilevel selection when connected with Eva Jablonka's 4 dimensional evolution. This means uniting the multilevel paradigm following from David Sloan Wilson and successors, with the four different "streams" of evolving entities, namely genetic, epigenetic, niche construction, and culture.

Within humans, I spend considerable time examining a topic that had been neglected by many evolutionary biologists since the selfish gene became the dominant ethos in that field, the study of human religious groups as units of evolutionary selection, which has gained traction in the last 6 years. I advance an even stronger hypothesis that some human religious groups are superorganisms, evolved multi-agent entities with specific properties and modalities above and beyond those needed to classify a group as an evolutionary unit. The consequences for the stability of altruism are varied and this hypothesis explains part of the altruism already extant among multiple religious communities. Considerable effort is turned towards clarifying and specifying religious superorganisms.

Moving from the birds eye view to the first person perspective I examine the specific psychological and cultural mechanisms through which cultural forces, including mythology early on, and religion more recently, reorganize individual psychology in altruism eliciting or stimulating ways.

Deliberate altruism at a collective level has been tried under different state and religious agents before, and a new form of it, more grounded in science, microeconomics, and ethical philosophy emerged with the formation of Effective Altruism, a movement I was an active participant in initiating and influencing, which we will occasionally visit during this examination of altruism in general. EA (Effective

Altruism) has a sui generis view of how to go about influencing the world in altruistic ways, so it will not be the bulk of our examination, but a perspective from which to look at altruistic entities as we go through the course of this writing.

Lastly I examine multiple social, political and global phenomena and draw ways in which they influence or are influenced by altruistic behavior, both extant and in potential.

What I hope to bring a new here includes:

- A new classification schema for bioculturally evolving entities, from genes to species and beyond,
- A perspective about altruism in the present and the future in the intersection of those examined in other disciplines in academia as well as by Effective Altruists (and their philosopher counterparts, utilitarians and consequentialists),
- A defense of the full blown case of human superorganisms, and within it a reclassification of some but not all religious groups, national groups, ethnic groups as belonging to the same biocultural supercategory.
- An examination of the scalability argument from multilevel selection theory that explores the frontier beyond Sloan Wilson's latest *This View of Life* (2019) and into the far future.
- An integration of the scientific consensus about altruism in different areas with this same far future oriented impartial perspective
- An integration of metaphor and analogy as conceived of by Lakoff and Hofstadter with the study of altruism at large, and altruism in relation to artificial systems in particular
- An extension of the discussions initiated by Derek Parfit (1986, 2008) on the relationship between personal identity, on the one hand, and altruism as well as consequentialism on the other, including concepts from anthropology, neuroscience, and artificial intelligence that were not present in Parfit's original discussions.
- A frequent conversation with the perspective of Effective Altruism
  An intellectual and practical movement oriented towards having individuals find what the most good they can possibly do with their lives is, and then do it.

Effective Altruism emerged from the conglomeration of far future oriented groups concerned about risks and benefits of human capability transcending artificial intelligence in Berkeley, and philosophers (professors and students alike) in Oxford who desired to leave the utilitarian, consequentialist armchair and go into the field and actually implement altruism maximizing strategies in the world. Throughout the second

half of this work, I will often refer to an `altruist` or an effective altruist having in mind someone with this mentality, or at least intention, of doing either the very most good they can, or at least someone set on a quantitative and very large set of positive consequences. MacAskill defines EA thusly (2019):

> Effective altruism is:
> (i) the use of evidence and careful reasoning to work out how to maximize the good with a given unit of resources, tentatively understanding 'the good' in impartial welfarist terms, and
> (ii) the use of the findings from (i) to try to improve the world.

In the first half however, I will be in dialogue with the current disciplinary work on altruism from each field's perspective.

I conclude the analysis with a novel yet encouraging *yes* to our initial question: given what we know about altruism currently in its extant and potential forms, there are stable equilibria which could both be reached from the point at which we currently stand across multiple dimensions, but also that could be sustained or enhanced in the long term up to cosmic scales of time.

What kind of altruist are we?

When you mean well, What do you mean? The altruism hydra.

Let us start off by saying that there have been numerous studies about many different aspects of altruism and that among those two broad classes can be distinguished: there are those who consider an act altruist based on the *motives* behind the act, and those who have *action based* definitions of it. To study someone's motives for an action requires two inferential steps: an observation, followed by an inference of motive. This inference can be innumerably complex, involving hidden causes and motives, beliefs that are never explicitly stated, and innumerous potential confounding factors (for a thorough analysis see Batson 2011). For this reason we will restrict ourselves here to action-based definitions.

Let us turn now to the way altruism is envisaged by different sciences and philosophies, while, to reiterate, considering solely altruism in terms of behaviors and their consequences. Although not all thinkers and researchers of a specific class see altruism the same way, we try to capture the overall tone of how altruism is seen in those areas.

The multiplicity of altruism(s)

At first I need to examine the inherently multifaceted nature of the concept of altruism. Common parlance uses the word "altruism" to mean several different things in different contexts - not necessarily consciously or explicitly. Those different definitions can be roughly divided into two broad classes, only one of which will interest us here:

- Intention/motives-based altruism, and
- Action-based altruism.

Here I will not address intention/motives-based altruism - such as when your intentions/motives are good, but the expected value of your actions, due to misinformation, is clearly negative. (e.g. if you think we should reinstate slavery worldwide for the greater good and take action in that direction).

Instead, I will focus on an action-based notion of altruism, that is, conceptions of altruism that are somewhat accessible, testable, and empirically verifiable in some way. I am not interested in the phenomenology of altruism, except in what Dennett calls the heterophenomenology (Dennett 1991) of altruism, as well as altruistic behavior per se. To mingle motives-based definitions with action-based ones causes both theoretical and practical confusion. This is one reason that I am not considering motives-based definitions in this piece. For an extensive treatment of motives-based definitions of altruism and related experiments, see (Batson 2011).

Even narrowing our focus to action-based altruism, we are still dealing with a Hydra, a many-faceted multivalent concept which different disciplines describe in different ways. In the original myth of the hydra, a multi-headed, long-necked beast possesses many monstrous, reptilian heads, and as soon as one is cut off, two new ones emerge. So to start with, we are faced with the Herculean task of taming the altruism hydra without letting the new muzzles spring out of our definition. I am aware that each of these fields doesn't use a monolithic definition of altruism for its discussions, and to some extent the names I give here are just guidelines to a principal component analysis that divides the concepts of altruism into discrete buckets. The labels to describe the buckets are a tool to facilitate understanding and represent compression, at the sacrifice of nuance. This type of problem is becoming commonplace in philosophy and in science, as our knowledge becomes more differentiated, nuanced and specific, while at the same time spilling between different buckets of conceptual closedness. But we have to start somewhere. And we will start with biologists.

Biologists

Biologists think of altruism in two ways.

1. *Strict biological definition:* considers an act altruistic if it generates more genetic benefit to genes an individual does not bear than to those it does. That is, if an act enables the replication of genes an animal does not have to a higher degree than it enables that of genes the animal has, including copies of those genes located in that animal's family members and community members.
2. *Loose biological definition:* an act is altruistic if it benefits other individuals more than it benefits the actor.

Kin selection and multi-level selection are subcases of 2, in which the benefit to other individuals is mathematically "justified" as it were by the genetic relatedness between individuals. That is, that individual's genes are still being beneficiaries, just not necessarily the copies that are inside their body or of their descendants, but those running in parallel through the evolutionary tree. This type of fitness which includes copies outside an individual is called *inclusive fitness*.

Social Psychologists

Social Psychologists envision actions according to the following table:

1. Altruistic - MaxOther
   Maximizes benefit - which already defined ex ante - to others
2. Cooperational - MaxSum
   Maximizes the total benefit including self and others.
3. Individualist - MaxOwn
   Only maximizes benefit to self, being neutral to outcome for others.
4. Equalitarian - MinDiff
   Minimizes the difference between any given two players, self included.
5. Competitive - MaxDiff
   Maximizes the advantage of Self over others
6. Aggressive - MinOther
   Minimize the benefit to others.

Game Theorists

Game theorists see altruistic actions as those whose outcomes benefit others at the cost of benefit to the agent.
$B(O) > 0 \wedge C(me)-B(me) > 0$
O = Other
C = Cost
B = Benefit

Moral Psychologists

Moral Psychologists - and Moral Philosophers of Neuroscience - consider actions altruistic in two distinct senses

1. Universal altruism: those would be cases in which a decision is made to take the action that maximizes overall good to all those influenced by the possible action palette.
2. Parochial altruism: frequently connected to Oxytocin (De Dreu 2012, De Dreu et al. 2011, De Dreu et al. 2015), parochial altruism is benefit to other individuals who share some identity, for instance belonging to the same group, clade, tribe, or team.

Different currents of psychology study states like happiness directly, as in positive psychology for example Lyubuomirsky and Seligman, or specific states such as gratitude and awe, e.g. Keltner and Shapiro. These states elicit or promote altruism, and occasionally prevent egotism as well, so they are causally entangled with altruism per se.

Moral Philosophers

Moral philosophers for the most part are not as interested in neuroscience as the ones I grouped with moral psychologists above, e.g. Guy Kahane, Joshua Greene, Julian Savulescu. Instead, most moral philosophers use other grounds, from abstract a priori reasoning, to divine command, through categorical imperatives and social norms to derive their conception of what Good is. Their conception of altruism varies accordingly. To a divine command theorist, helping someone else better follow the word of Allah could count as an altruistic action. Even serving God directly could count, as God is a different agent from, say, a devout catholic. There are many schools within metaethics,

including those that are rule based, deontology, those that are virtue based, virtue ethics, and, pluralist schools, which derive their ethical worldview from a combination of the precepts of other schools. In my philosophy education in undergrad I was mostly a consequentialist, transitioning during the masters in which I wrote a book, and in the years hence to a pluralist with a strong component of utilitarian consequentialism, and a smaller but significant component of virtue ethics.

Measurability is very complex within moral philosophy under non-consequentialist metaethics, so here we will focus on consequentialist or consequentialist leaning pluralist moral philosophies, because most points about altruism can be more easily made with access to quantitative reasoning. There are ways of translating moral ethics that are less quantitative into more quantitative ones (eg MacAskill 2020) so in focusing on consequentialist and utilitarian ethics henceforth I am not abandoning or de-privileging other ethical theories, many, if not most of the insights about altruism I will present will have an analogous version under the auspices of a Kantian imperative or Aristotelian virtue. I must however choose one "main" framework of attack, and that will be a broadly consequentialist view of metaethics, with utilitarian leanings but without specifying narrower questions such as which type of utilitarianism, which are less relevant for this body of work.

Effective Altruists and Consequentialists

Effective Altruists and Moral Philosophers of consequentialist inclinations consider an action altruistic to the extent it maximizes overall good. Within that framework, different individuals will have different perspectives on what the good is e.g. happiness, joy, wellbeing, PERMA[1], QALY[2]. Different philosophical currents also differ in who can be counted as a moral patient, that is, a potential recipient of goodness.

Most people consider present and future people to be moral recipients. Some extend this further; for instance, should animal happiness and suffering be qualified as part of "overall good"? Different philosophical currents also differ in who can be counted as a moral patient, that is, a potential recipient of goodness. Should cyborgs? Computer programs? Video Game characters (Tomasik 2017)? Fictional characters? Minds in the Matrix (Chalmers 2003)? People who do not exist yet? Fetuses whose brain is not formed yet, but whose body already exists? Atoms (Tegmark 2014 - Friendly Artificial

[1] PERMA, in positive psychology, is a measure of wellbeing based off the five main things people seem to want to accomplish to experience a positive life: Positive Experience,Engagement , Relationships, Meaning, and Accomplishments.

[2]Quality Adjusted Life Years, a frequently utilized public health measure of the value of life, often used to decide between different interventions that consume scarce resources.

Intelligence: the Physics Challenge)? Aliens? People in other possible worlds? People in the past?

These questions are responded to differently by different consequentialist philosophers and effective altruists, that is, who counts as a moral patient varies among people who have the same type of orientation philosophically and ethically.

During this body of work I will occasionally refer to what an altruist should do in some situation, or consider. In those specific usages, specially towards the second half of the writing I will be referring to what, to the best of my ability seems like it would maximize the amount good done by that agent in that particular situation, which you can consider to be an Effective Altruist way of looking at it, or just a goodness maximizer.

Everyone else

Common parlance uses the word 'altruistic' in many if not all the meanings above, as well as some others we will not consider to be altruism here, such as when your intentions are good, but the expected value of your actions, due to misinformation, is clearly negative (Wilson 2015). 'Altruism' is also often used interchangeably with philanthropy, which we can describe as altruistic giving. Keep in mind, as explained above, all these are *only the definitions that are action based, and expected value related*.

Motivation and Intention: Often altruism is conceptualized in terms not of the effects, or the actions taken by an agent, but instead by their intention or motive (Batson 1991;2011). We are not considering motives based definitions in this work: it would clutter our brains with confounding factors and definitions, of which there are too many, and motives are mere proximate mechanisms that only tend to evolve and persist in virtue of the altruistic habits that they (sometimes) contribute to. So they are causally redundant from the standpoint of fitness pressures as well as moral calculi. Altruistic acts can also be thought of as ultimate causes.

Multi-Buckethood

In his *Behave*, (Sapolsky 2017) neuroscientist Robert Sapolsky tackles the problem of multi buckethood head on:

> First you can't begin to understand things like aggression, competition, cooperation, and empathy without biology. I say this for the benefit of a certain breed of social scientist who finds biology to be irrelevant and a bit ideologically suspect when thinking about human social behavior. But just as important, second, you're just as much up the creek if you rely only on

> biology; this is said for the benefit of a style of molecular fundamentalist who believes that the social sciences are destined to be consumed by "real" science. [...]
>
> How are we supposed to make sense of all these factors in thinking about behavior? We tend to use a certain cognitive strategy when dealing with complex, multifaceted phenomena, in that we break down those separate facets into categories, into buckets of explanation. [...]
>
> Putting facts into nice cleanly demarcated buckets of explanation has its advantages - for example, it can help you remember facts better. But it can wreak havoc on your ability to think about those facts. This is because the boundaries between different categories are often arbitrary, but once some arbitrary boundary exists, we forget that it is arbitrary and get way too impressed with its importance. [...] In other words, when you think categorically, you have trouble seeing how similar or different two things are. If you pay lots of attention to where boundaries are, you pay less attention to complete pictures.

With Sapolsky, I hold that there are no physical correspondents to these disciplinary buckets, as each action is the product of all biological, physical and social influences that preceded it and will be causally relevant to all that come after it. We cannot say that a behavior is caused by a single gene, culture, episteme, hormone, trauma, Schelling point, weltanschauung or decision theory. We may attempt to draw a function which describes a distribution of causal efficacy between different levels, but restricting the analysis to just one bucket would thwart any serious attempt at a causal analysis.

Moving from ultimate to more proximate, I examine the existing definitions and theories concerning action-based altruism according to their respective disciplinary origin, beginning by entering the realm of biology, and deepening the analysis in the realm of anthropology.

Altruism in Biology

Definitions

Altruism

Biologists think of altruism in two ways. The strict evolutionary definition is that an act is altruistic if it generates more genetic benefit to genes the agent does not bear than to those it does, that is, if an act enables the replication of genes an agent does not have in higher degree than it enables that of genes the agent does have, including copies of those genes located in that agent's family and community members. The same gene, that is, can be present in an individual genome, and in a relative (Hamilton 1964). The strict definition does not differentiate which copy is having its fitness increased by the actor.

The loose biological definition contends that an act is altruistic if it benefits other individuals more than it benefits the actor.

Ultimately this distinction is a matter of temporal scope. An interaction where an animal sacrifices a seeming indicator of fitness may turn out to be a longer term investment - for instance in reputation, or chances of future mating - if the resolution lens of observation has a wider temporal scope.

Benefit

Benefit, in biology, has different definitions as well. In the case of genes, a benefit is the existence of more copies of that gene, either over time or across space. In the case of an individual, that is less clear. Arguably, a biologist could say a benefit is what makes an organism more fit. That just kicks the can down to reproductive fitness, which itself is cached out into a vocabulary that usually involves whether the genes you carry manage or not to replicate, and, in addition, whether the epigenetic structures of that individual get to be passed onto the next generation. Reproductive fitness is an arbitrary concept, though, as it requires the preemptive choice of a spatiotemporal range, within which one can look at environments and their organisms, and tell which ones seem fit for that environment based off on their long term (within the spatiotemporal range) genetic fitness.

A benefit in biology can thus be thought of as the fraction gain of inclusive fitness from an event. This inclusive fitness is parcelled as, in vast majority, genetic, and secondarily, composed of other heritable structures subject to natural selection, such as epigenetic structures in a manner analogous to the genetic one in kin selection.

Otters and Others: *cui bono* altruism in biology

The philosopher Daniel Dennett dabbles in many sciences, and provides a clear description of an important question to ask, to biologists, anthropologists and philosophers alike, from *Darwin Dangerous Idea (pg 324):*

"Lawyers ask, in Latin, *Cui bono*?, a question that often strikes at the heart of important issues: Who benefits from this matter? The same issue arises in evolutionary theory, where the counterpart of Wilson's actual dictum would be: "What's good for the body is good for the genes and vice versa." By and large, biologists would agree, this must be true. The fate of a body and the fate of its genes are tightly linked. But they are not perfectly coincident.
What about those cases when push comes to shove, and the interests of the body (long life, happiness, comfort, etc.) conflict with the interests of the Genes?\

But when push comes to shove, what's good for the genes determines what the future will hold. They are, after all, the replicators whose varying prospects in the self-replication competitions set the whole process of evolution in motion, and keep it in motion.
[...]
Natural selection is not a force that "acts" at one level—for instance, the molecular level as opposed to the population level or organism level. Natural selection occurs because a sum of events, of all sorts and sizes, has a particular statistically describable outcome. "

Biologists have settled into two arbitrary boundaries that inform their concept of altruism, the *gene* and the *individual*.

The Individual

Individual

When it comes to individuality, in biology it is not always considered to be the monolithic four dimensional individual spread across time with assumed identity of 100% (Lewis 1983) with themselves - as is the case in the partially jocosely nicknamed homo economicus, for instance. It is, instead, often a fractional identity based on the genetic similarity between the individual taking the action and the individual benefiting from it. The famous line attributed to Haldane about sacrificing oneself to save 3 brothers or 9 cousins (which secures more genetic benefit than sacrifice) entails that there can even be more fitness of an individual outside the individual than inside it, that is, more of an organism's alleles can be preserved in subsequent generation if that organism dies than if it lives, conditional on that providing great benefit to other organisms that share some of those alleles. Biologists often slide imperceptibly back and forth between the notion of an individual as one continuous entity spread in time and space, the 4 dimensional object we usually tend to normally think of when thinking of a person, or an animal, and this fitness-based, genetic notion of individuality, which I will call genetic sliver individuality.

This genetic sliver individuality isn't related with divisibility as the word originally intended but instead it is just a name for a fraction relationship of how much of the constituents of the genetic composition of an animal is still extant in the population later.

Biological altruism in the fitness sense is not too likely to sacrifice individuality, as the fitness-related conception of individual is not deeply related to what we call individuals normally (see discussion below). The loose definition, however, sometimes involves some loss of individuality which is offloaded into the other biological systems that are carrying the unit of selection (frequently the gene) forward. So the loose

organism definition of biological individual sets the stage for kin selection and other mechanisms of replication via others to partially deteriorate individuality.

The individual is usually taken to be the boundary determining self and other. Individuals are usually constituted of parts that share a common *telos* (Deacon 2011), which is sometimes described as a striving, as an urge to linger on, to self propagate, as a bundle of constraints or as a structure composed of goals, such as a utility function. Even though we have extended phenotypes (Dawkins 1989), parts of our phenotypes that are not in our bodies, and extended minds (Chalmers & Clarke 1998) part of our mentality that is not in our body, the physical boundary of our skin is for most purposes clear enough to determine where the individual ends and where the world begins.

Despite its apparent simplicity, the individual is far from a "settled concept." (Clarke 2013, Ainslie 2003) - In-dividual was a bad name choice, we should probably have taken the clue from the physicists when they named the atom a-tom, 'cannot divide', and later divided it. We are slow learners, as a people.

Besides some of the conceptual difficulties to classify individuals for the purposes of kin selection that we examined so far, the notion is also problematic in a plethora of other realms, scientific or not, let us look at some of them in turn:

In the different branches of social and psychological sciences the individual has often also been constructed as mosaic, disjointed or constituted only in relations to the other. In Canaque society - a population living not far from Australia - for instance, the individual isn't the fundamental unit of identity, what constitutes the Canaque identity is first and foremost the position of a phenomenologically centered world within a sequence of events, and instead of a person being the unit which experiences the event, the fundamental entity is a substructure of a 4 generational cycle. So that one is the "same" representational person as one's great grandfather in an ever repeating cycle.

The notion of groups (Nurit Bird-David 2017) can also be variegated across different societies which see univocal multiplicities, or multiplicities of one, ones as mereological constituents of a group, or as separate entities from the mereological class to which they belong.

The Cartesian subject of "I Think, I Am" (which was badly translated in English as I think therefore I am) is also thought of as a phenomenological continuous string in time.

Not only in Anthropology and philosophy does the individual thread an unstable line between Scylla and Caribdis, in politics we see a similar phenomenon to the Canaque society 4 generational with a striking resemblance to the more recent book *The Fourth Turning* which has become an important influence in the decision process of the White House's chief strategist, Steve Bannon, who, contradictorily enough, tends to be obsessed about the individual in his movies (Bannon 2016). In the political arena and

in the social sciences it is the mascot of libertarians who praise the concept as a foundation to determine the organization of their proposed political system. It is the enemy of collectivists, who want some collective to be the unit of most relevance in the constitution of political entities. Interestingly, bioanthropology and biogeography unveiled a psychological correlation with rice crop farming and large levels of collectivism, specially studied in China, where genetic predispositions to individualist and risk taking behavior have been partially selected out from the population in areas with rice crop farming but not in wheat areas, most famously the DRD4 7r allele (Leung et al 2005).

The divide between the individual and the group, when does one end and the other begin, has been recently studied by philosopher Ellen Clarke (2014; 2016) in the context of selective pressures and evolutionary dynamics as well as by anthropologist Terrence Deacon in the context of the emergence of semiosis. We will now turn our attention to this divide. This is specially important to our general question about the scalability of altruistic systems in the future because some types of synergies and interactions are only possible between or within individuals, or between and within groups.

Ellen Clarke's model of individuality

Ellen Clarke is an Oxford philosopher researching the nature of individuality, both in abstract ontically substrate independent, and in our earthly biological systems. Her work can help us understand the process of altruistic agent composites becoming individuals as well as individuals decomposing into agents.

The type of issue examined by Clarke can be exemplified by this question:

'Why did not natural selection, acting on entities at the lower level, disrupt integration at the higher level?' (Maynard Smith and Szathmáry 1997, p 8)

Let us begin with her definition of evolutionary transition:

"I (Clarke) define an evolutionary transition as a shift in the hierarchical level at which heritable fitness variance is expressed in a population, before distinguishing different kinds of question that can be posed with respect to transitions and then setting out the conceptual ingredients required to answer each kind."

Reproductive autonomy also plays a role in our conception of transition:

'Entities that were capable of independent replication before the transition can replicate only as part of a larger whole after it' (Maynard Smith and Szathmáry 1997, p 8)

On a compositional or mereological view, what unifies the different transitions is

the process of wholes becoming parts that are physically nested inside new higher-level wholes.

Since Ellen Clarke has written several papers on the transition between group status and individual status her work is particularly relevant for us here. In order to tackle the question of individuality, she uses ontologically neutral terms (Quine 1950) which help conceptually clarify the desiderata. She describes a difference between a separate individual - a state one object - like a single celled organism, and a state two object, like a polar bear:

> A State One population is divided into objects with the following properties – they are living, and they exhibit heritable variance in fitness (figure 1). To understand the second claim, assign each object in the population a character value or trait, z, and a fitness, w. The population must meet the following three conditions: i. Character value (z) must vary so that different members of the population express different values for z (phenotypic variation). ii. Variation in z must correlate with variation in fitness, w (differential fitness). iii. Fitness (w) must be heritable (fitness is heritable). Together these conditions guarantee that the State One population is capable of undergoing evolution by natural selection (Lewontin 1970). The objects in a State One population therefore compete against each other in a standard one-level selection process. We call the objects in State One population 'organisms' or sometimes 'biological individuals'. A State Two population is identical except that the objects into which it is divided, and which exhibit heritable variance in fitness, are themselves aggregates of former State One objects (figure 2).3 These aggregate creatures can also be called 'organisms', but their parts cannot. For precision, I add the further constraint that State Two populations exhibit zero heritable variance in fitness at the level of the parts (as in Gardner and Grafen's 2009 'superorganism'). In other words, heritable fitness variance is exhibited exclusively at one hierarchical level, in each case. (Clarke 2014)

In most of the transition literature (Szathmáry 1995), the main question asked is what makes group level changes stable over time. That is an interesting question. But here I am more interested in the question of whether there are intermediate phases that are relevant for a concept of individuality relevant to our types of altruism. This is a complex question in the intersection of science and philosophy.

Since this is complex, it might be worth illustrating with an example borrowed from analytic philosophy: The concept of health is determined by a two fold process. First, we arbitrarily decide what situations and parameters would count as being healthy. Then, using that arbitrary concept, we find scientific instruments and theories that help us determine if someone is or is not healthy, such as, say, blood tests. So to assess if an individual is or not healthy we go through these steps. Individuality is constituted in a similar way

The philosophical aspect is to establish what do we mean when we think of individuality. The scientific aspect is, conditional on that pre-established concept, to try to understand which real world entities satisfy the conditions.

It may also turn out to be the case that the thing that determines the relevant kind of individuality - to distinguish organisms from groups and superorganisms - itself is an

empirical datum, in which case it is a composite of two questions, one purely scientific, and one in the intersection of science and philosophy. This would be the case if Individuality is like health, that is, if it cannot be a purely abstract concept floating in the proverbial void.

By using the terms 'state one' and 'state two', Ellen creates a neutral conceptual landscape that is individuality agnostic, which is necessary for conceptual clarification (Quine 1950).

The transition of state one to state two is not without friction. At every level inside complicated groups of organisms that experience some group level selection, there are still entities whose incentives are to act against the interests of the larger whole of which they are part. Let us look at them in turn:

- Individuals have incentives to act against groups, by doing things like parasitism, predating on common goods, cheating, or implementing any other self benefiting strategy that isn't prosocial.
- Individuals have incentives to act against groups and groups might develop counter-strategies to prevent individual defection, for instance the scapegoating mechanism (Girard 2007): a complex regulatory mechanism that prevents ultimate violence - war - by iterating smaller forms of violence - the creation and destruction of real or symbolic scapegoats.
- The group level can also tackle the lower levels (individual included) by for example (DeScioli 2013) creating a "moral" cognitive system based on third party dynamic coordination to reduce friction to the group at a cost both to individuals and to morality. If Descioli is correct, our moral sense when it comes to punishment was actually designed not to follow any actual moral guidelines, but instead to just automatically side with one side in context such that less conflict would emerge as a result of disagreement or fight between two parties in a larger group.
- Within cells there are also battles between the incentives of different parts, possibly the most famous of which being the one that lead to the emergence of sexual reproduction, where female organelles defeated male organelles, becoming the sole providers of mitochondria and other subcellular components for the next generation (Ridley 1994), this later had enormous effects and is considered to be the first domino in a cascade that led to the great majority of secondary sexual characteristics that distinguish females and males, due to differential parental investment (Trivers 1973).

It is in this context that we can think of evolutionary battles not between

organisms, but between levels - such as the group, individual, subindividual, supergroups etc... and the tension between the individual and other intra-individual and extra-individual levels of selection becomes more clear. The process of evolution is ever unfolding, and in it, the individual is constantly fighting against lower and higher levels of complexity. The evolutionary battleground isn't merely constituted by a pandemonium of minuscule replicators - Dawkins's minimal unit of selection - in a free for all battle who only serve themselves of larger structures for their own benefit. It is instead extant also in the dimension of levels of resolution, which themselves are fighting against each other in mathematically describable ways (e.g. Price equation and derivatives). This orthogonal dimension of evolutionary friction is the locus of the tension between individuals and groups, and individuals and their subcomponents (Dennett 2017, Simler 2013).

These intra-level battles are of utmost importance to our organizing question of the scalability of altruism. It is only if the multilevel confluence of factors that elicits and sustains altruistic behavior and components continues to propagate itself through the aeons that we can expect larger and larger, or deeper and deeper, structures of cooperation and altruism to permeate the future of our species and our technological, biological and cultural descendants.

The Gene

Why is the gene selfish?

"Be warned that if you wish as I do, to build a society in which individuals cooperate generously and unselfishly towards a common good, you can expect little help from biological nature. Let us try to teach generosity and altruism, because we are born selfish."
— Richard Dawkins, *The Selfish Gene*

If you want to transmit a technique for making arrows to your descendants' descendants (Mesouldi 2011), it is impractical to pantomime it and expect that the recipient of the information will be attentive to the details of your behavior that matter (Tomasello et al. 2005) "how sharp it is" but not attentive to the details that don't matter "whether it needs being created in cloudy days, or while kneeling". Behaviorally, humans are dispositionally prone to over-imitation, also called superstitious imitation or procedural imitation. Differently from Chimps (Tomasello 2009; 2005), who seemingly only copy actions that are perceivably causally necessary to achieve a goal, we are prone to copy the entire ritual leading to a goal instead of going straight for the prize. In one

experiment where a box was open by people through a sequence of steps, some ritual an unnecessary for the box to open, and some causally necessary for it to open, chimps ignored the ritual gestures and copied only the necessary steps to open it, while human children copied the steps religiously, even when causally unnecessary. Two possible reasons why evolution would equip us with that heuristic: 1) Biological processes often require some level of redundancy to achieve robustness (Minsky 2007) and 2) and more relevant for our specific discussion is that the goals of human actions are frequently all too complex to be understood immediately by someone who learns it and is unable to keep track of all the causal tree roots leading to the desired outcome (Mikhail 2007). Planting and sustaining crops, or following elephants to find water during the long dries of the desert of Mali are complex behaviors which involve a "leap of faith" on the part of junior learners, learning it cannot be accomplished by understanding the full causal structure of the relevant events, but instead must involve anticipating a future, or an intention, which precludes them from being learned by just copying what is visibly causally relevant (Henrich 2017). No wonder we are more superstitious than the other animals. This behavior is an incipient form of superstition, which cumulatively develops into ritual, and, in our species, into full blown religions.

An easier alternative to guarantee fidelity over generational time is to have a representational template, such as written language, to transmit the information, this is tantamount to transforming a fluid behavior into a structured, discrete description of it, and transmitting the description instead of the behavior. Musical notation as well as written language are tools enabling this sort of higher fidelity transmission (Blackmore 2000).

The same process took place billions of years ago in biology: instead of the behavior and proteins contained inside an organism having to pass by direct transmission of the proteins and behavioral dispositions themselves, evolution stumbled on a medium of information storage that was more discrete, more robust, and more representational than the preceding loose state of affairs, namely the transmission of information via codes of sequences of nitrogenated basis, which progressively increased the amount of template information storage and decreased the amount of direct dynamic information storage (where the information is stored on the thing itself). This gave rise to what we call genetic code composing today's RNA, DNA and the coding elements of viruses. Evolution learned how to write, which enabled it to play Chinese whispers better. Transmitting information with higher fidelity through digital compression.

This compression enabled a massive scaling of altruism in the past in its biological form, since inclusive fitness is derived from indifference between a particular DNA structure and a high fidelity copy of it, so any altruism towards kin that relies on inclusive fitness is in some sense dependant on this compression discrete strategy. Looking through our organizing question of scalability, we can see we are currently in

another process of great scaling caused by the compression of information in our written languages, print languages, and now digital languages, including computations over mathematical digits, all of which have enabled a literal global scaling of our economies, and thus, of the entangling of destinies of distant organisms. Although this process has not been without cost, and has made the superstructure of global economics brittle in some parts, it also strengthened and gave huge robustness to other parts, and overall has enabled levels of altruism between people hitherto undreamt of.

Why at the Gene's Level?

> *"The claim that the gene-centrist perspective is best, or most important, is not a claim about the importance of molecular biology, but about something more abstract: about which level does the most explanatory work under most conditions. Philosophers of biology have paid more close attention, and made more substantive contributions, to the analysis of this issue than to any other in evolutionary theory."* - Dennett - *Darwin Dangerous Idea*

The issue quoted above in Dennett suggests another one: Why did the level of resolution for static storage that evolution found turn out to be the gene level, and not something smaller, or larger?

This is related to the gene level being the most robust template made of discrete parts[3] (ATCG U) able to store enough information to encode a characteristic where natural selection can exert force. The discreteness of nitrogenated bases permits discrete storage, and genes are the shortest code sequences able to code for a fitness constrained characteristic.

Genes store most of the information because:

- the mistakes are predictably interesting: The probability of a valuable mistake in the copying process of DNA is high compared to analogous combination, which tends towards the mean for instance in a one dimensional characteristic.
- the representation is digital and

[3] Parts have to be discrete or close enough to discrete because if for instance a child simply "blended" the characteristics of both parents, being half and half, the strength of the attractor returning everyone to the mean would be too strong. Furthermore, characteristics that emerged via mutations and were beneficial would dilute over time in large enough populations. Sorting that which natural selection exerts force on in discrete parts allows for cumulative evolution, as well as redistribution of genes in sexual living beings. Biological entities cluster continuous data into discrete units(colors, sound tones, mouth sounds, neuron activation)

- It permits activation and deactivation of clusters, an IF THEN clause written in biological code: A gene can operate as a program that calls some, but not other, genes into functioning, both for a lifetime and conditional on some specific activation mechanism.

It is easier for genes to be selfish - evolutionarily optimized for their own preservation and replication - than for any other upper level structure, up to the organism level. Let us unpack this statement: If you imagine time as a left to right X dimension on a two dimensional graph, there will be more robust and persistent objects, which are present for a long span of length to the right of where they first emerged, and there will be objects that persist for less time.

The same is true of structures, for instance, Theseus' ship which replaces a piece at a time, but remains with the same structure throughout. Now what Darwinian evolution is all about is finding out how to stretch objects the most extension in the time axis, and there's two ways for stuff to accomplish this feat.

Genes and diamonds: ways to everlast

Genes and diamonds represent two distant sides on a spectrum of "how to be robust over time": while diamonds persevere by being fairly resistant to external disruption, genes persevere by creating the conditions of homeostasis in cells of the germline such that copies of themselves get sequentially passed to each new cell generation. The diamond versus gene strategy is somewhat analogous to the plants versus animals. While plants evolve hardness, resilience, and poison because they cannot move, animals evolved the skill of fleeing, hiding and moving around to avoid local harm.

The diamond perseveres via hardness and, except human made diamonds, it was not optimized for the task, it just accidentally turned out to be hard. On the other hand, the gene is optimized to persevere via survival of the fittest.

The fittests what? Genes, alleles, individuals

Although the main beneficiaries, the recipients of *cui bono* in biology are genes, the easiest boundary to determine behavior is the individual. We can more easily distinguish "who is acting?" than "which genes are causally responsible for this action taking place." Thus the mathematics of altruism (Mc Elreath & Henrich 2007) in biology are organized at the individual level. This leads to calculating the benefit of an action to an individual by summing the benefit the individual herself accrues with the benefit

received by other individuals in proportion to their relatedness (the expected percentage of shared genetic material between them, e.g. ~50% for a sister and ~12.5% for a cousin[4] [5]).

Maybe to confuse non-biologists, geneticists use the term 'allele' to refer to what evolutionary biologists and popular science books call 'genes'. The difference is that a gene to the geneticist represents a *location* in the genome which is occupied by one among different variants of alleles, each of which encoding different information about which proteins to assemble and when. So what an evolutionary biologist would call a "gene for smooth peas" is, to a geneticist, "an allele for smoothness of the gene SBEI, which also has another allele for wrinkles". Red hair is coded for in the M1CR genetic locus, and at least three different alleles code for different variants of pheomelanin production.

Gene's eye view; how many me-copies out there?

What changes when we view altruism from the Gene's eye view? What changes is that the set of actions considered egoistic is significantly expanded, and insignificantly contracted. That is, there are a lot of actions that help copies of your genes that happen to not be located in your body, but not that many actions that help you as an individual but do not also help the genes contained in your body. So the set of actions considered selfish expands.

Which gives us a notion of altruism based on actions that benefit individuals who lack that set of genes, any other action is deemed selfish - that is the gene's eye view. In other words, divide the biological world into "genes I have" and "genes I don't". Anything that benefits the ones I don't more than the ones I do is considered altruistic - within this biological framing. A gene is a template, which accrues evolutionary benefit only in cases where the template is replicated and re-instantiated in later generations. Unlike diamonds it was not optimized to make sure that it itself persists several generations from now, but instead that its structure is replicated and passed on intact to future organisms that are alive then, often relatives.

Altruism was for long considered a human specific domain until the morality of primates and other mammals began to be studied. Under different conditions, chimps, elephants and even slime mold amoeboids exhibit behaviors that we recognize as altruistic (Sapolsky 2017; DeWaal 2006).

---

[4] If you were ever confused by the mystery of us sharing 50% of genes with our siblings but 98%+ with Chimps the solution to the mystery is that when talking about heredity between humans we only count the genes that vary between humans and ignore the invariant ones.

[5] This expected relatedness measure ignores some particularities such as the genetic material in mitochondria, the size difference between the Y and X chromosomes and other details.

We will revisit the gene's eye view of altruism in the sections below on *kin selection* and *group selection*.

Biological theories of altruism and cooperation

Biological research has mostly addressed not the causes of altruism, but the conditions under which it becomes stable -- that is, an evolutionarily stable strategy for at least some population members, through the course of evolutionary time. That is because it was previously believed by earlier genetic paradigm biologists such as Hamilton, E. O. Wilson and Trivers that a group of altruists *always* can be successfully invaded by a group of cheaters, and if cheating evolved in a group of altruists, it would quickly grow in population. This stood unchallenged until (Dennett 1995) demonstrated how evolution's search algorithms managed to find some niches that are actually altruistic, when new replicators emerged whose incentive structure to preserve fitness operated differently from the structure of genes, and thus, who could work in tandem with genes to constitute altruistic individuals. At the same time, in biology proper, a revolution was happening which would lead to multilevel selection ascending to dominance in the field, with David Sloan Wilson as its strongest advocate (Lewens 2015).

Biology also addresses related concepts -- synergy and cooperation -- which appeared earlier in evolutionary time compared to full-blown altruism. Let us look at them in turn.

Cooperation and Synergy

Cooperation is an action requiring agents. It is a subset of a concept that does not necessarily require agency — synergy. Synergic interactions happen, for instance, when two chemical compounds facilitate the production of a catalyst for each other, or in other ways facilitate the continuation of the chemical processes and constraints that generate or maintain each other. Synergy arises from the intrinsic limitations of autocatalysis and self-assembly processes, and leads potentially to the evolution of complex systems (Deacon 2011; Corning 1998).

As such synergy led to the evolution of organisms capable of agency, some of them established control systems which allowed them to act in a different manner towards the organisms with which they are synergistic. Simple organisms, such as groups of amoebae *Dictyostelium Discoideum*, for example, are able to make the "choice" over whether to cooperate with another group of amoebae or not. If previous interactions had the other group allocating more amoebae into a reproducing sub-group

than their "fair share," a group will store this information and interact with the previous one less frequently. The amoebal equivalent of losing - or gaining - trust (Sapolsky 2017).

This means that the incentive alignment or misalignment between two biological entities exerts strong selective pressure on organisms. Amoebae are among the simplest living systems, and even then the evolutionary process found a way to optimize their ability to cooperate, defect, and do so at better odds than chance. The biological entities in question in this case is a group of amoebae, not individuals, but since they share genetic material at very high levels, they have what could be considered genetic sliver individuality, though not physical individuality.

The scaffold for cooperation available to amoebae, however, is very brittle compared with the many ways in which animals, in particular brained animals, can cooperate. Animals have not only genetic sliver individuality, but also well-determined physical boundaries, or physical individuality, which allows evolution to develop plentiful strategies for cooperation. The five best known ones, as described by Martin Nowak (2005), are as follows:

Kin selection

Animals must rely on indirect measures to anticipate - unconsciously - the expected value of cooperating or acting altruistically towards another individual. Hamilton summarized the shape of a payoff landscape that incentivizes evolution to maintain that kind of kin-selected altruism.

$$R > c/b$$

The level of - *genetic* - relatedness to the beneficiary of one's action (r) must exceed the cost to benefit ratio for the individual (c/b) in order for cooperation to be favored.

Although for pragmatic purposes this is thought of only in terms of genetic relatedness in the majority of cases, the same logic would apply to a relatedness calculus that involved both the genetic lineage and the epigenetic lineage of an individual, proportional to the heredity of that epigenetic factor, that is, including epigenetic factors only to the extent that they are hereditary. In as much as epigenetic inheritance is heritable in the same way genes are, we should expect that their relatedness is the same.

Historically, kin selection was considered both the first and most powerful theory for altruism in biology -- it seemed to rest on solid ground. Yet, in 2010, it took a hit in the article *The Evolution of Eusociality* (Nowak, M., Tarnita, C. & Wilson, E. 2010). which dismantled its importance, with data showing how the development of eusociality in certain animals did not follow the inclusive fitness theory, and proposed instead that the

causative agent leading to higher cooperation between closely related individuals is instead the formation and the persistence of groups, in which an evolutionary transition ends up occurring.

Direct reciprocity

Direct reciprocity, abstracted from Trivers, shifts us from relatedness in space to relatedness in time. Whereas kin selection requires for evolutionary stability that the relatedness supersede the cost (c) benefit (b) ratio, direct reciprocity requires that the expected number of interactions in the future exceeds that same ratio. So if (w) is the expected number of future interactions within a dyad,

$w > c/b$

A caveat to this strategy is that it only works if there is some uncertainty as to the number of interactions, as defecting on the last round is always the optimal strategy if the machinery in which you are implemented anticipates it is the last round. The components of w are uncertainty times each individual potential future interaction.

Indirect reciprocity

Nowak (2006):

> "Direct reciprocity relies on repeated encounters between the same two individuals, and both individuals must be able to provide help, which is less costly for the donor than it is beneficial for the recipient. But often the interactions among humans are asymmetric and fleeting. One person is in a position to help another, but there is no possibility for a direct reciprocation. We help strangers who are in need. We donate to charities that do not donate to us. Direct reciprocity is like a barter economy based on the immediate exchange of goods, whereas indirect reciprocity resembles the invention of money. The money that fuels the engines of indirect reciprocity is reputation."

Indirect reciprocity is most pronounced among humans, who possess several cognitive abilities that enable, facilitate, and maintain it, such as the ability to represent symbolically, language and gossip. It also has participated in the evolution of human-specific abilities such as morality and social norms.

Among humans, indirect reciprocity often involves trust or reputation, there are two ways of knowing if an individual is trustworthy: through knowing him, or through knowing his reputation. If the proportion of reputational false positives and reputational false negatives is the same, it doesn't really matter to choose whether to cooperate, if you

know the person or the reputation. This means that for simplified systems, many-agents interaction happening in discrete swapping dyadic interactions will entail the same consequences for an agent as dyadic interaction with the same person.

If moving from kin selection to direct reciprocity we moved from twins in space to twins in time, we are now dealing with virtual twins. Indirect reciprocity can be maintained if the probability (q) of knowing someone's reputation exceeds the cost benefit ratio of the altruistic act.

$q > c/b$

Network selection

Network selection involves choosing who your interaction partners are. In most animal species proximity is a precondition for interaction. Although in an evenly distributed group cooperative strategies are not stable, if groups of cooperators can assort themselves together, cooperation becomes a stable trait in that part of the network and is not invaded by defectors. It has been speculated that high trust societies, and within larger societies, high trust neighborhoods, are the human equivalent of that process.

For network selection to stabilize, the average number of interactors that any given agent has must be less than the cost to benefit ratio. This initial treatment leads to the catch acronym of the three R's: reputation, reciprocation and retribution (e.g. Trivers 1971; Alexander 1987; Haley & Fessler 2005; Nowak & Sigmund 2005).

A similar conceptual construct in economics is the theory of the firm (Coase 1937; Williamson & Winter 1993) which contends that a firm's size will be proportional to the size that minimizes internal and external transaction costs - that is a firm will expand in number of people until it becomes taxing to do so.

Group Selection

Group selection is somewhat rare, and the logic behind conditions for its evolutionary stability becomes more mathematically complex as they involve more quantities and details.

If selection is weak, and groups seldom split, *then* a mathematical formalism equivalent to those above arises. For a max group size (n), and (m) groups, we have:

$c/b > 1 + (n/m)$

Emerging as a sufficient condition for the stabilization of cooperation in the long run. I discuss group selection at length below.

Of the 5 strategies described above, only kin selection and direct reciprocity involve unproblematic notions of individuality. In the case of indirect reciprocity and indirect selection, it is assumed that the individual benefiting is fully the same over time, including, for instance, fitness, which actually changes over time as aging corrodes the body, or positively when for instance status changes increase reproductive fitness.

There are also conditions for altruism that were not discussed by Nowak in his *Five Rules for the Evolution of Cooperation.*

Relaxation of Selection

thx

All the constraints and mechanisms explored so far are conditions for the *stability* or *maintenance* of altruistic conditions. Hui and Deacon (2010) propose a mechanism instead for the *emergence* of altruism in humans, for how it came to be: the relaxation of natural selection. To think of relaxation, I can illustrate with a metaphor. Imagine a fitness landscape (Sewall Wright 1932), where peaks are higher fitness. Selection pressures can act as a waterline, "drowning" those who aren't fit enough for long enough. Relaxation of selection is the lowering of that waterline.

How does that type of relaxation process potentially incentivize altruistic interactions? This can happen if there is relaxed selection for autonomy - for instance two different redundant ways of obtaining that which can be obtained via autonomy, one of which not autonomous but other-dependant - The presence of others and of altruistic others offloads some of the functionality leaving room to degradation or modification of that phenotypic expression into other functionalities, or none.

Relaxation is one avenue for the creation of altruism because it is by nature a degenerative process, and the degeneration of autonomy creates the prosocial conditions for emergence of altruism. Relaxation can also enable differentiation into complementary subsets. A process that might have led to conditions that enable humans to be a remarkably altruistic species, where the relaxed pressure on some autonomous functionality differentiates into other fitness enhancing functions which take for granted the presence of others, locking in prosocial adaptations, and enabling altruism. For example partially losing the ability to, say, hunt, autonomously creates the conditions of interdependency that allow for, or enable, new synergies to be created between hunters cooperating, and thus opening a gate to altruism between the individuals. The

functional codependence between these individuals creates the very conditions for functional differentiation and emergence of altruism. Even the unconscious calculus of when to be altruistic could be a result of lack of selection for precise calculus of cost benefit of cooperation, in a manner analogous to the complexification and decrystallization of the song of the Bengalese finch post domestication. A cognitive adaptation for cooperating under narrow circumstances could have been relaxed into a cognitive adaptation that acts fully altruistically, to benefit others, as a result of degeneration of that narrow calculus. This could be the result of changes into the social dynamics of groups for a variety of reasons, from increase in prosociality to higher density, physical proximity, excess clumped resources among others. Any of these could elicit the prosocial conditions which would relax the pressure to cooperate precisely and open the possibility of behaving altruistically, the emergence of altruism.

Relaxation of selection can sometimes be crucial in the process of differentiation of individuals and groups across evolutionary time. If you relax the constraints of survival enough via social facilitation from co-specimens, over time the individual loses the capability of being alone, effectively becoming a social individual. In the extreme cases, loss of autonomy includes loss of reproductive autonomy - such as in eusocial species - loss of autonomy may also lead to division of labor within groups. The emergence of eukaryotic cells involved actual fusion of individuality into one being, and loss of autonomy of it's previous constituent parts. The birth of sex had similar effects. The most impressive case of constitution of individuality arising from loss of autonomy was the transition from unicellular to pluricellular life forms - contended to have happened a staggering 25 times! (Clarke 2014) In all these cases, there was loss of autonomy, although whether there was relaxed selection remains to be determined.

Drift

Genetic drift is the movement of the gene pool that happens not due to constraining forces such as natural or sexual selection, but instead simply due to the chips falling where they may. This process may compound over time in statistically unlikely cases, and thus drive populations and their gene pool to move in some direction. Frequently, a founder effect, or the separation of a population from the larger population with which it interacted, has some statistical deviance from the totality of the gene pool that gets preserved through many generations.

Genetic drift can result in deleterious or neutral mutations, and thus it can result in mutations that increase probability of altruistic behavior even when these do not coincide with natural or sexual selection forces. Relaxation of selection can speed up the potential maximum speed of drift.

Error

Evolution frequently designs unreasoned reasoners (Dennett 1995), organisms that act according to rationales, but with no explicit representation of their own reasons. as well as heuristics (Eva & Norman 2005, Twerski 1974) that shortcut and round expected benefit calculations in the cognitive systems of animals. This makes them fail in predictable and interesting ways (Ariely 2008), some of which we could call *accidental altruism*. This can be used to produce a taxonomy:

Accidental Altruism: Predictably irrational cognitive action that leads to altruistic consequence in a majority of cases.

Erratic Altruism: Predictable behavior, not necessarily cognitive, which leads to altruistic outcomes, including in simple animals, or even protozoa.

It may bear pointing out that manipulation (Dawkins 1999) of another animal is, from the perspective of the manipulated animal, altruism, falling under the scope of erratic altruism if it is an exploitative systematic strategy.

Failure Altruism: Individual cases of altruistic behavior that do not seem to be part of any larger scale pattern.

Clumped Resources

Sometimes cooperative and altruistic strategies are needed because resources are clumped together (Kropotkin 1902, Wynne-Edwards 1962). In economics/philanthropy the idea of social capital has come to attention recently, and it relates to the inverse effect. When health damages sustained by an individual are larger than the resources they have to tackle it, altruism also might be necessitated.

Clumped resources increase incentives for individuals to develop trustworthiness (Hui & Deacon 2010). Trustworthiness requires a model of whether an individual is stable over time, and, from the perspective of the individual, to see themselves as individuals and the same over time facilitates abiding by this tacit agreement. Thus, clumped resources facilitate altruism as well as facilitate individuality.

Transgenerational Altruism

Mostly via niche construction, some animals end up creating benefits for several generations of not necessarily biologically related beings. Religious scribes who didn't reproduce in the middle ages are a human example (David Sloan Wilson 2010), in this case, the prosocial values promoted by a religion - parochial or not - are transmitted through time by individuals who forgo direct genetic transmission via celibacy.

Matriphagy, or eating the mother, happens in some species of spiders, in this case the mother herself becomes the niche of growth of her children, giving the energy stored in her body to them even more than a mammal mother does. Coral reefs also partake in a similar process of creating valuable niche assets to future coral reefs, including their own descendants.

Altruism Among Humans

We begin with two somewhat contradictory perspectives from Darwin:

"It is extremely doubtful whether the offspring of the more sympathetic and benevolent parents, or of those who were the most faithful to their comrades, would be reared in greater numbers than the children of selfish and treacherous parents belonging to the same tribe. He who was ready to sacrifice his life, as many a savage has been, rather than betray his comrades, would often leave no offspring to inherit his noble nature."
— Charles Darwin, Descent of Man ... p. 146.
[Sympathy] will have been increased through natural selection; for those communities, which included the greatest number of the most sympathetic members, would flourish best, and rear the greatest number of offspring. - Darwin 1871, p. 130.

The scope of possible cooperative and altruistic interactions between agents has been increasing nearly monotonically since the emergence of pluricellular organisms: cells whose destinies covary gave place to organisms whose destinies covary with their families. But it really took off with groups of humans and the beginning of contemporary history. After the agricultural revolution, interpersonal correlations enabling cooperation became geographically broader, and by the industrial revolution, there was already some correlation between the destinies of farmers in China and traders in Liverpool, despite the monumental geographical and cultural barriers that separate them (Hobsbawn 2010). With the rise of global markets and stock exchanges, the world economy - the world's ability to allocate resources and quantity of resources to allocate - became heavily interdependent. The formation of an economic and cultural elite in the West that spread its values worldwide gave this further momentum. In the last blink of an eye, by historical standards, the internet led to instantaneous opportunities to cooperate and defect with people thousands of miles away. Social media, in particular Facebook, Twitter, Instagram, Snapchat and TikTok increased the social stakes of online bets by a significant amount. Reputations became permanent tattoos (Enriquez 2013), and symbolic identities became reified as online identities, especially among the wealthiest billion of people.

Through most of that history, there was an alignment between personal benefit to increase in cooperation. The very existence of non-rival goods in a social species with a truthful language inevitably led to the spreading of ideas, with stronger fixation of good ideas (Mesoudi 2011, Diamond 2011). Although many wars occurred, people respond to incentives, and the near monotonic increase in the benefits of cooperating across vast distances helped in the process that led to a decline in violence over nearly any time scale (Pinker 2011).

Let us look at some human-specific perspectives on the nature of cooperation and altruism. Starting with our cognition which is substantially more complex than other animals, to a point where we have developed a moral cognition, a set of tools through which we process information about moral questions. A few of these are shared with large apes at least in homologues, but they are substantially stronger in humans.

The Automaticity of Altruism

Before we examine the preconditions of moral cognition, it is worth mentioning that human altruism has features like momentum, derived from the proclivity of one's sociocultural milieu to encourage/discourage prosocial behavior (Keltner et al 2017). Further in some form of prosocial experiments, the longer subjects are given to make a decision, the less altruistic they become, indicating that prosociality is to some extent automatic. It can also be stimulated by compassion, touch and oxytocin, to which we will return later. (Hertenstein et al 2009; Crockett et al 2008). Humans often give 40-50% to strangers in resource sharing experiments, even in small scale societies (Henrich et al 2005). This impulsive altruism might seem baffling to the game theorist at first, but on reflection it isn't. For starters behaving in a different way in experimental setting than in our day to day life is computationally costly and can be socially costly (what if you just think no one is watching, or that there are no consequences, but it turns out you were being watched) so it makes sense that we are automatically nice to strangers even in contexts where there is no explicit benefit to ourselves. Gratitude triggered by previous prosocial action also leads to more frequent prosocial behavior (Gordon et al 2012). As Keltner et al put it, "cooperation is a relatively intuitive snap judgment" whereas self interest requires a more cognitive perspective taking.

Besides being automatic, prosociality is also contagious (Keltner et al 2017; Nowak and Roch 2007; Schnall et al 2010).

The Preconditions for Moral Cognition

To bridge the gap separating the moral psychologist, moral neuroscientist, and evolutionary psychologist's models of moral cognition from the AI scientist's modus operandi in agent design, we need first to understand and then formalize some processes

on which moral cognition depends, i.e., its cognitive preconditions:

(a) *symbolic acquisition*, which permits representing an absent agent or situation by using a symbol that stands in its place;

(b) *analogical reasoning*, which permits relating different situations and transferring moral weight between them in proportion to how analogous they are; Analogical reasoning is often composed of metaphorical cognition, which I shall examine as well.

(c) *Preference Inference*, which allows us to probabilistically guess what others want or "want to want' (Smith et al 1989); and (d) *Theory of Mind*, which allows us to conceive of others as moral agents like us, with minds of their own and wants of their own.

(d) *Theory of Mind*, which allows us to interpret other moral patients as having a mind and experiences of their own. Together, these seem to me to be the main preconditions of moral reasoning.

Let us zoom in and look at each in turn:

(a) **Symbolic acquisition** is a complex process whose development has been extensively studied by Deacon (1998). The process of symbolic acquisition involves a process with three distinguishable modes of reference, each necessitating its predecessor. Our species' unique ability to manipulate *symbols* has formal precedents in the acquisitions of *icons*, which refer by similarity or indistinguishability; *indexes*, which refer by contiguity or correlation; and *symbols*, which still refer while being functionally severed from the original semantic network connections through which they were internalized, thereby becoming amenable to the structural and recursive processing familiar to psycholinguists and computer scientists. The resulting symbolic cognition is needed to reason and communicate about counterfactuals and remote situations, which play a role in moral reasoning (*e.g.* Gärdenfors 2007, Tomasello 2005).

(b) **Analogy** is the core of cognition, contends Hofstadter (2003). However, given that analogy carries a superficial appearance of non-formalizable smoothness, this statement is rejected by many computer science researchers who attempt to work around it. I agree with Hofstadter despite not giving his "core" emphasis too much credit; a model of cognition is bound to fail if it tries to ignore how we use correspondence and partial similarity to project knowledge from a domain to a target domain with which it shares some formal structure. Within analogies, metaphors stand out (Lakoff 1980) as being even more informative to moral cognition, in virtue of intrinsic asymmetries and developmental features. This paragraph alone boasts 23 metaphors spotted by the authors' first count, and uncountably many analogies.

Analogical and **metaphorical thinking** are better understood in humans when thought in relation to linguistic cognition. Which metaphors we use to understand our

relations, including moral and romantic relationships can actually determine whether they go south (Lakoff 2015 personal conversation) or not. I use "metaphor" here in a somewhat technical sense developed in cognitive science and linguistics over the last 45 years by Jaines, Johnson, Lakoff (1980),.

For example, conceptualizing a romantic relationship as a third person, or a voyage, or a road, or a container, might cause or prevent conflict. To understand the human-specific metaphors, and in particular, the human-specific *primary* metaphors (Lakoff 2015 personal conversation) - which are ontogenetically prior - would be necessary to understand how our moral system works.

Let us look at another example: The concept of moral credit and debit, where a container is imagined to bear the banking properties of morality, permitting one to deposit good actions, and withdraw immoral ones while remaining a moral person by being a moral creditor to this imaginary bank. This can be used as a foundation to calculate fair trade with a romantic partner. Differently, having a sacred foundation which believes some actions are impermissible in a sacred manner - say violating the adultery commandment of the Hebrew bible - might give one a different metaphorical frame in which to reason about mores. A fight could ensue where within the frames established by their cognitive metaphors, both parties believe to have the moral high ground.

Cognitive Metaphor, a special type of analogy: It seems we reason very often through Cognitive Metaphor, as described in the works of Lakoff (1996) and Johnson (Lakoff & Johnson 1980, 1997; Johnson 1994). As (Gentner et al 2001) contend in their namesake article, metaphor is like analogy, so many of the constraints which determine necessary conditions for analogical reasoning will also be operating forces here. Cognitive metaphor theory however has an extra layer above its core which is particularly relevant to moral reasoning; it distinguishes primary from complex metaphors. Metaphors that are developmentally and logically prior, and all others. Restricting ourselves to moral cognition, primary metaphors can roughly be thought of as cognitive programs that can be triggered during a developmental window, which determine via which metaphors an individual will come to think about moral concepts. Like a duckling can be imprinted to follow a human or a dog if that is the first animal it encounters upon hatching (Lorenz 1937), a human can be imprinted to thinking about morality as care or as obedience to authority depending on her family structure during crucial periods of development. There is a distinction between metaphors and analogies. Metaphors are more specific than analogical reasoning in crucial aspects such as the trigger time-window, predefined asymmetries between source and target domain, and sometimes, learning inevitability, thus they are more formative of our moral cognition.

(c) **Preference Inference** is the capacity to abstract an agent's preferences or goals from behavior, (Lucas et al 2014) summarizes a few experiments in the child learning literature, combining the Luce-Shepard choice rule with a Bayesian model into a Mixed Multinomial Logit model seems to provide an account of preference understanding

among young children. A related idea, preference elicitation, can use Markov decision processes to infer preferences in concept-space conditional on previous known preferences of the same agent (Rothkoph & Dimitrakakis 2011). Preference understanding is relevant for morality because preference satisfaction is an important aspect to be considered in moral choice, with choices that satisfy more preferences being often assigned higher moral rank (utilitarians and specially preference utilitarians see them as crucial, either instrumentally or terminally as moral goals). Whereas theory of mind studies have mostly focused on understanding others' mental beliefs, inferring preferences and goals plays an important role in our ability to understand others and is more fundamental to moral reasoning than belief inference.

(d) **Theory of Mind** is the capacity to conceive of others as having a mind or mental states and acting on them. (Schaafsma et al 2015) critically surveyed the field, and laid out a deconstruction and reconstruction multi-stage formal division of the process, attempting to maintain a parallelism with the cognitive neuroscience data which is precisely what I am trying to lay out here with most psychological concepts examined. Theory of Mind matters because morality prescribes how to treat others, which is arguably unachievable without understanding what confers them moral patienthood, often considered a derivative of how they *feel* and their degree of autonomy. A model of autonomy and experiential dimensions in agents, with dimensions such as magnitude of pain, magnitude of pleasure, and magnitude of meaning often complementing otherwise descriptive conceptual models of one's immediate external reality (in ethology: ümwelt) and action pattern drives. That is, in addition to representing a combined model of the world and feedback systems from higher cortical layers adjusting and shifting these models to fit our predictions, we also experience the world, and being in the world (dasein, numenon, Ümwelt) in ways that need to be understood to capture all dimensions considered by our moral cognition.

Aspects of Moral Cognition

Moral concept acquisition: Concept acquisition is a mountainous task, a subset of which is moral concept acquisition. The beginnings of a theory of agent independent cognitive conceptual acquisition can be picked out in Sotala (2015a; 2015b). A moral concept depends on many levels of abstraction from simpler concepts combined which can be represented in vector spaces according to *Conceptual Spaces* (Gärdenfors 2004) - the notion of "fairness" for instance cannot be abstracted directly from "give" or "take", but is mediated by abstract categories such as "trade", "exchange", "credit", "debit", "distribution" and "symmetry" among others - it is also mediated by functors in concept-space (Mikolov, Yih & Zweig 2013) which slide meanings in parallel towards a predictable direction, as a homogeneous force field.

Where and When in the Brain Moral Cognition Happens

In reasoning about morality so far, I have focused on structural analysis of the preconditions for our type of moral cognition, and the functional description of moral concept acquisition in a logical space. Now I will emphasize the specific, and possibly contingent, physicality of how moral cognition is processed in our brain. Many people sneeze when gazing at light with the corner of their eyes, not because there's an evolutionary function for it, but because the bundles of nerves downstream of light rod activation in some eye regions, and the bundle of axons responsible for motor activation of the sneeze reflex happen to be near one another. A short circuit like effect activates the sneeze not for functional reasons but due to accidental proximity between these bundles. Likewise, aspects of our moral cognition should be expected to be the product of evolutionary tinkering, and understanding how to locate where in our minds moral reasoning takes place might help us alleviate, de-bias, and predict moral behavior and moral choices.

Moral Development

We are not born with a fully functional moral cognition equipment, Moral cognition develops through stages in which we acquire the preconditions above as well as aspects of moral reasoning itself.

Moral Development: which capacities become available to us at which stages of development during an individual's life history.

From Preconditions to Altruism per se

Having a general sense of the preconditions and processes involved in creating our moral capacities from a cognitive standpoint, we can move on to altruism and cooperation per se, through a myriad of lenses. Let us begin by wearing our social psychologist goggles:

Social Psychologists

Social Psychologists envision actions according to the following table:

1. Altruistic - MaxOther

Agent Alt maximizes the payoff that others receive.

2. Cooperational - MaxSum

Agent Coo maximizes the collective payoff. Self included.

3. Individualist - MaxOwn

Agent Ind maximizes its own payoff.

4. Equalitarian - MinDiff

Agent Equ minimizes the payoff difference between players.

5. Competitive - MaxDiff

Agent Com maximizes the difference between own payoff and that of others.

6. Aggressive - MinOther

Agent Agg minimizes the payoff received by others.

Both Coo and Alt agents are prosocial in this case.

An Ind can be prosocial in case of goal alignment between it and those it is interacting with, so agent Ayn Rand can shake hands on a mutually profitable deal.

An Equ individual can be prosocial if there is alignment between their interests and that of the players they are interacting with. In small settings between individuals, not groups, equalitarianism is often practiced (sharing chocolate or a dinner bill equally). Equalitarianism can be, in small settings, a mechanism for prevention of conflict under simplified information. As illustrated by this comic strip:

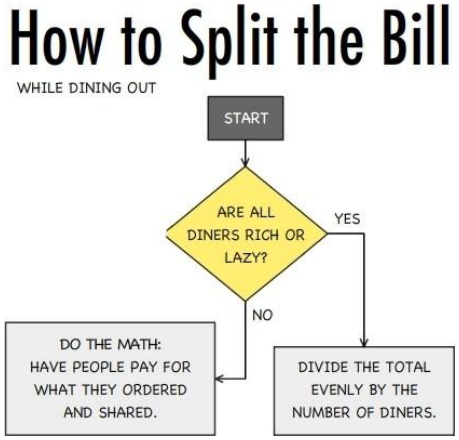

Fig 1.0: Splitting bill flowchart

Rich and lazy are the two conditions where the cost of processing information is higher than the cost of being equalitarian. Since humans erroneously believe that equal division is "fair," despite humans being bundles of properties that differ enormously quantitatively and qualitatively, minimizing difference prevents conflict in some settings.

All of the definitions of agents above use the usual monolithic agent baseline definition of individual in order to compute which kind of policy an agent is following.

Moral theorists

"Morality binds people into groups. It gives us tribalism, it gives us genocide, war, and politics. But it also gives us heroism, altruism, and sainthood." — Jonathan Haidt

Moral Psychologists and Moral Philosophers with an interest in neuroscience consider actions altruistic in two distinct senses:

1. Universal altruism: those would be cases where the brain decided to take the action that maximizes overall good to all those influenced by the possible action palette.
2. Parochial altruism: frequently connected to Oxytocin (Kret & De Dreu 2013, 2016), parochial altruism is beneficial to other individuals who share some identity, for instance belonging to the same group, religion, clade, tribe, or team.

   Notably, parochial altruism can be primed to make one class (e.g. female) more salient than another (Asian) (Sapolsky 2017) whereby people favor more their cohorts in the primed dimension.

Effective Altruists and Consequentialists

Effective Altruists, a movement springing from moral philosophy and from safety research in AI, and moral philosophers of consequentialist inclinations consider an action altruistic to the extent it maximizes overall good (MacAskill 2015). Within that framework, different individuals will have different perspectives on what the good is e.g. happiness, joy, well-being, PERMA[6], QALY[7], SWB[3]. Effective Altruism differs from

[6] PERMA, in positive psychology, is a measure of well-being based on the four main things people seem to want to accomplish to experience a positive life: Positive Experience, Relationships, Meaning, and Accomplishments.

[7] Quality Adjusted Life Years, a frequently utilized public health measure of the value of life, often used to decide between different interventions that consume scarce resources.

utilitarian philosophy despite most Effective Altruists being of utilitarian inclination. Effective altruism is a movement, a philosophy of how to live - by trying to maximize the amount of good you cause through your life - and not an ethical position or moral argument. Although it has a philosophical element, it is a philosophy of action, and more action than philosophy. There are also Effective Altruists who are pluralists in ethics or who have other metaethical positions, but still agree with the actions and maximizing principle of doing the most good one can do.

What they have in common though is the approach of maximizing the aggregative expected value over some dimension. This notion differs starkly from parochial altruism in our contemporary world as for most individuals in developed countries, the biggest altruistic opportunities tend to be donating either far away in time, e.g. ripple effects, or far away in space, e.g. schistosomiasis reduction in sub-saharan African countries.

One for All and All for None: self sacrifice in game theory

Are cooperators altruists and altruists cooperators?

That depends on which discipline you ask. Social psychologists have a metric for

[3]The tripartite model of subjective well-being, a concept developed by Ed Diener which is a measure of how people experience the quality of their lives and includes both emotional reactions and cognitive judgements.

social value orientation which distinguishes cooperators from altruists, as seen below:

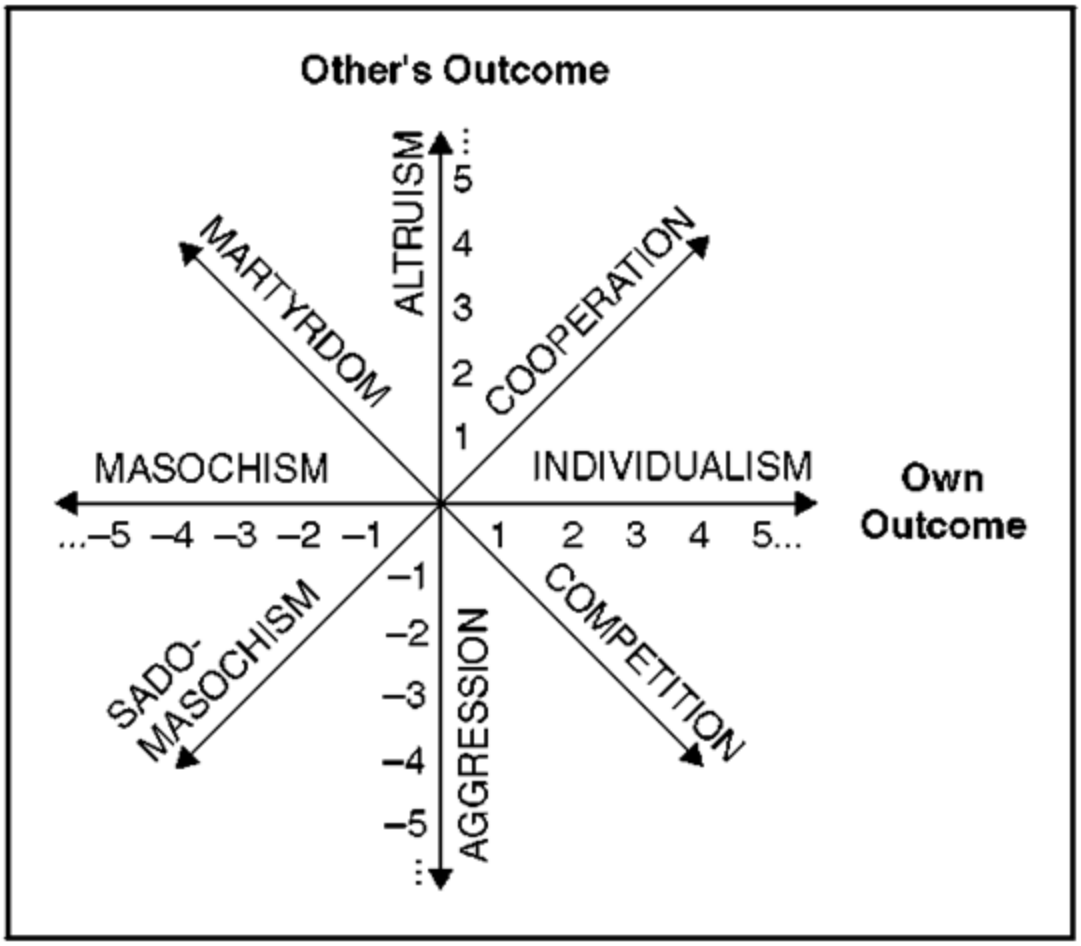

Fig2: 8 axis of agency
adapted from *The Fire Burns Within*, Cambridge, 2018;

Within social psychology and game theory, Max-sum counts as cooperational behavior, caring about yourself as much as about others makes you a cooperator, not an altruist. It makes you somewhat altruist, and somewhat selfish.

Philosophy

Altruism in philosophy is usually thought of as pure other benefiting, with a notable special case being other maximizing. The tradition is influenced by Von Neumann-Morgenstein's utility theorem which, conditional on some assumptions about what rational behavior is, suggests that a rational agent will maximize expected value.

Expected value doesn't necessarily have a pre-established referent and it could mean value to self or others, in the altruistic case, others. But it is not always the case that benefiting others even at the expense of self is the most effective way of benefiting others.

Steve Omohundro (2008:2012) and Stuart Armstrong (2013) have shown in the context of flexible or self modifying agents, including seed AIs - AIs with a potential to become AGIs (Artificial General Intelligences) - that the dichotomy between other-maximizing and self-preservation isn't a dichotomy simpliciter in real world scenarios. That is because any entity with goals that require manipulating the world in the future has incentives not only to take direct actions towards those goals, but also to preserve the existence of such goals themselves, either in the entity or in others. Not all entities are smart enough to realize that self-preservation and goal-structure stability are instrumental goals to nearly any other goal it could have, but we should expect that some evolved systems, such as animals, and some artificial systems, if sufficiently capable of abstraction and world mapping (Armstrong 2013), will notice that self-preservation is an instrumental goal to most altruistic goals. In abstract or in local scenarios it may be easy to distinguish actions that benefit self versus others, but in more complex environments with more complex agents, some boundary trade-offs begin to appear.

Most evolved creatures avoid self-sacrificing, and when it was in the interest of evolution to force them into death, this was done via molecular clock changes accelerating aging far more often than installing a conscious module creating Freud's hypothesized death instinct. A wish for self destruction is seldom an effective instrumental goal, though low status people, most often women, sometimes commit altruistic suicide, psychologically triggered by a self perception of low value, and hypothesised in humans to save community and family resources. In males a killing spree or other forms of risk taking violence are more conducive to low quality mates becoming attractive to some women. Notably school shooters often get love letters and sex letters after partaking in their murders. Males aren't attracted to women who perform these violent acts, so females commit suicide instead, incresing resources to their relatives.

Tversky and Ainslie Contra Parfit

Unlike Anthropologists, Philosophers for the most part ignore such difficulties and discuss the abstract question of benefiting self versus other as equivalent to selfishness versus altruism.

We philosophers do however discuss the question of whether future selves have identity with one's present self *simpliciter* or if there should be some form of discount in

how much a self-interested agent ought to be willing to receive rewards in future time. Conversely, philosopher Derek Parfit frequently (e.g. 2010,1984,1971) turned this question on it's head asking: to the extent that we are willing to collaborate with and invest in those future beings that share our name, some memories and some psychological characteristics, should we not also consider helping others in the present? If there is no further question of personal identity once you carve into those similarities, then, he would claim, this excessive prioritizing of future selves is a *bias*, not a feature. This is in stark and direct apparent conflict with the picoeconomics notion of hyperbolic discount (Ainslie 2003) and in general with the literature on cognitive biases that people discount time steeply (Kahneman Tversky 2000) so it bears some explaining. The mismatch seems to me due to using two notions of value interchangeably that are not interchangeable.

One is the interpersonal, market, societal, economic, value of something. This one we discount more strongly than would be rational under the assumption that we are the exact same agent over time - that we are monoliths. Our discount is both steeper than would make economic sense, and also it is hyperbolically shaped, meaning that an external agent (say a broker robot) programmed with exponential time discount could trick us into losing value over time, by offering us to purchase rewards that are in the proximal future, and buying from us what would have been cumulatively more rewarding in the distant future, as do most casinos and social media platforms.

The other notion of value is the subjective value to the individual. This can be far more than the economic value to society. And for those who dissolved the question of personal identity in philosophy, it sometimes may also be a very different question, since the notion of individual value changes when there is no individual, or when the individual isn't stable over time. I propose two ways of considering the dilemma faced by Parfitian philosophers, on the one hand one could choose a different dimension that isn't time in which to do discounting. Many people choose to discount over distance in mindspace, which roughly translates to how similar or dissimilar the personality and mind that would be receiving this reward is from mine determines how much I'm willing to pay for it. So what in economics is assumed of time (distance in time tracks value) in this case will be tracked by similarity in mindspace.

Alternatively, one could embrace a notion of discontinuous discount in time that is very steep, lasting however long a thought, impression, or emotion might last. Philosophically well informed immediatism of self seems to me consistent with many, if not most, conceptions about philosophy of mind and psychology of self. Some traditions such as mindfulness, vipassana, Dharma and Tao, endorse taking this view at least some of the time. To me the interesting question that emerges from this view is: given nearly all the evolved design of our cognitive architecture, as well as the cultural design of our mind's software, is predicated on there being at least some level of agent-continuity over time, even if you hold this view philosophically, most of the time your mind will be

driven to take actions that are related to a future more than 2 seconds ahead. And given that is the case, but you don't consider those people you, doesn't that make Parfit's argument for altruism even stronger? In general it seems to me that Parfit's argument for altruism is as strong as one's discount, in any dimension, is steep. The fewer *me's* there are, the more reason I have to dedicate to *others*. This is perhaps most clear emotionally to people who are told they will die, who are less likely to go the "bucket list" way, and more likely to want to finish their deeds.

Value then is used as a multi-meaning concept (Block 2003 calls it a *mongrel concept,* but the term didn't catch) it sometimes means individual value - which itself requires decomposing "individual" - and sometimes means interpersonal, economic, market, societal, inter-temporal or marginal, value. That creates the perceived contradiction between philosophical views and psychological or economic views, but the contradiction disappears if we call one $value_1$ and other $value_2$. That is both notions of value are meaningful, but they refer to different things through the same word.

By contrast to social psychology and game theory, remember that from the perspective of utilitarians or effective altruists, since you are looking into a sample of all minds when choosing who to benefit, the self becomes a vanishingly small fraction of those you can help, and therefore there is little reason to distinguish "maximize all" from "maximize others", in practice. Once your circle of empathy expands enough, maximal cooperators and maximal altruists behave a lot alike.

In that sense, utilitarians and Effective Altruists can be useful 'tools' to bring about or sustain states of altruistic equilibrium that would not come about through the customary structures of value and their ensuing incentive vector fields. The existence of entities trying to maximize overall good, regardless of to whom, enables bridges from current states to unlikely states, as a form of directed stochastic hill climbing algorithm. To the extent that our original question of scalability depends on moral cognition, values, and incentives shifting, it is very useful to have a fraction of people, however small, engaged in a lifelong project of causing as much good as they are able.

Return to Biology: The Selfish Allele Making the Altruist Individual

Alleles (the molecular biologist word for what less-specialized areas call genes) under normal conditions optimize for there being more copies of themselves in the future. This happens regardless of whether it is that physical instantiation that is present in the future, a copy will do. Even a defective copy whose milieu contains a corrective ribosome which will later make it a perfect copy will do.

Copies of alleles are spread over space, individuals, groups, species and *time,* but fitness only coerces their distribution, in the sense Dennett showed above, in the time dimension and the quantity dimension.

In the long run alleles don't thrive if they are doing better than their neighbors, they thrive if they are doing better than the average allele. A token (instantiation) of an allele that codes for cancer, multiplying itself uncontrollably, could produce many copies, but if the mutation that gave rise to it only happened in somatic cells (that do not go through the germline), it would be in for a surprise[8]. One reason why biologists say natural selection is short-sighted.

For the same reasons, an allele that codes for individual-selfish behavior in a species in which more altruist groups tend to outlive more egoistic ones will peter out. The allele for individual-selfishness, and the selfish individual, may seem to be doing well from a Darwinian perspective, if compared to their neighbors. Yet, with high probability, their group might end up dead, leaving the gene pool. Altruism can win in this case not because there is a new level of causation that reverses reductionism, and applies downward causation which originates in groups. Altruism thrives because the average long term fitness of each allele that coded for it was higher than that of alleles that code for individual-selfish behavior. Wilson (2010), the main proponent of multilevel selection theory, drawing from his own work with Elliott Sober, uses this strange but informative definition of altruism stemming from multilevel selection theory: altruism in terms of relative, not absolute, fitness. A behavior is selfish when it increases the fitness of the actor, relative to other members of its group. A behavior is altruistic when it increases the fitness of the group, relative to other groups, and decreases the relative fitness of the actor within the group.

This leads us to a notion of altruism which privileges groups over other levels of resolution, including the individual. In this case, there is a clear tension between altruism and individuality, as the promotion of groups as the level at which selection happens offloads part of the fitness constraining characteristics from individuals to groups of which they are members, and in cases where redundancy can be removed or deteriorated, it opens a door for individuals whose constitution is less autonomous evolutionarily.

For a visually satisfactory and more thorough dynamic explanation of the evolution of trust and distrust, the website https://ncase.me/trust/ has incredible short games that provide the basics of mathematical trust and cooperation evolution including a flexible game (on bubble 7) where one can twist knobs of types of players and observe the iterated dynamics in ways impossible to describe in written form and acquire the cognitive intuition for. Depictions of population dynamics games over periods of time

[8] The exception being transmissible cancers, observed in Tasmanian demons.

with different constraints and types of players. The *paragon of the polygons* does a similar analysis with regards to segregation and outcomes. https://ncase.me/polygons/ Both are outstanding work at transferring knowledge to intuition without which I'm not sure I would have understood these concepts or altruism itself as well as I have during this research process.

We have evolved incredibly sophisticated and complex game theory processing cognition to deal with all these scenarios while continuing to pass on our genes and memes. Our contemporary notions of Good and Bad are a projection from these forces operating over time, they are abstract concepts that compress millions of years of information obtained through the iterated games played over the eons into simple categories so we can manipulate them in our minds and use them to think and make decisions. But how did they start? How did the egg of a neutral world, turned into a full chicken able to cross roads for good and bad reasons, via playing millions of games of chicken?

To that question we will turn now.

The Emergence of Good and Bad

An animal's Ümwelt (Uexküll 1931), in ethology is a concept that evolved in part from Kant's notion of Noumenal world, and from the computational theory of mind and action-based behavioral frames, which suggests that an animal will have a map of the world that organizes itself around the functions that animal needs to execute in the world, not necessarily reflecting the lowest resolution computational accuracy that its cognitive system is capable of processing. Neurons that fire together, wire together. Conversely, neurons that need not fire, often transition into different functions, both developmentally, and, once adults, through plasticity and synaptic culling. This leads animals to often have a functional representational structure of the world, a non physical one. We see doors as portals, cups as containers, and street lights as levels of permissibility. Even though we see little of the world as us, we see a lot of the world as what it is *for* us.

Reasoning about the emergence of semiosis, Deacon contends that the process of reference making, of something meaning something to something, starts alongside the beginning of life and individuation. He creates a notion of "ententionality" (Deacon 2011) as a wider scoped biological form of the philosopher of mind's intentionality, and because wider, with a larger extension. The assumption, there, is that what constitutes individuation is a cluster of constraints, or, maybe more precisely: an attractor state over a process in a cluster of constraints that becomes responsible for the lingering on of the same process with its constraints. That is a process becomes a generator of itself, while

using mostly self generating chemical and energetic resources. Once a system is such that the processes that secure its continuity are embedded in it, is now has evolutionary incentives to begin distinguishing what is good or bad *for* it. So a protozoan may detect some molecule and react by moving towards or away from it: the origins of "good" and "bad."

Later down the evolutionary road, it may begin to differentiate further good into "food," "indication of nearby desirable genetic material," "gradient that indicates food in that direction," and other non explicit representations, all embedded in the way that the constraints process the "invasion" by an alien entity. So if this analysis is correct, the differentiation of the individual and the world happens at the juncture where there start to be things that the individuals can interact with *as* they are *for* themselves, that is how they change the ways those constraints - constitutive of the individual - are or are not able to sustain themselves over time.

A cluster of constraints that self-preserves and is energetically stable might begin to have "reasons" that are good for it, or bad for it. And those will be instantiated in its behavior, at least as a compression of the innumerous possibilities of what a foreign object in contact with one's body could be. So semiotically a plethora of meaning like the one our minds are able to entertain starts phylogenetically as simplified undifferentiated meaning. Not so simplified as to defy the necessary discontinuity which characterizes individuation, but simplified no less. Furthermore, in this perspective, facts and values start off together, and only much later, when each fact can be used differently to take different actions with different goals, the need for a reasons-neutral map begins to emerge. Accuracy comes long after pragmatism. In semiosis as in the saying: *Better wrong than dead*.

The Individual, in Biology and Anthropology

The individual is usually taken to be the boundary determining self and non-self. Individuals are usually constituted of parts that share a *telos* (Deacon 2011), which is sometimes described as a striving, as an urge to linger on, to self propagate, as a bundle of constraints or as a structure composed of goals, such as a utility function. Even though we have extended phenotypes (Dawkins 1982) and extended minds (Chalmers & Clark 1998) the physical boundary of our skin is usually clear enough to determine where the individual ends and where the world begins, a few complex boundary cases notwithstanding.

Despite its apparent simplicity, the individual is far from a "settled concept" in the political arena and in the social sciences. In politics it is the mascot of libertarians who praise the concept as the foundation to determine the organization of their proposed

political system. It is the enemy of collectivists, who want some collective to be the unit of most relevance in the constitution of political entities.

Anthropological studies of conceptualizations of the individual also examine the difference between populations. Among some populations, the individual, or, to use a less loaded label, the human, has also been constructed as mosaic, disjointed or constituted only in relations to the other by different branches of social and psychological sciences. In Canaque society - a population living not far from Australia - for instance, the individual isn't the fundamental unit of existence, what constitutes the Canaque identity is first and foremost the position of a phenomenologically centered world within a sequence of events, and instead of a person being the unit which experiences the event, the fundamental entity is a substructure of a 4 generational cycle, so that one is the "same" representational person as one's great grandfather in an ever repeating cycle.

A weaker, but far more studied distinction is that of individualistic versus collectivist societies. Collectivistic societies, particularly those that descend from rice crop cultures, such as most of mainland China, incentivize less individualistic behavior, and people are more attentive to their relative position in a network instead of seeing the world as centered in themselves. Collectivistic cultures have a few distinct genetic characteristics such as a lower frequency of the r7 allele of the dopamine receptor D4 gene (DRD4) (Sapolsky 2017) and a significantly higher frequency of the short (S) allele of the serotonin transporter functional polymorphism (5-HTTLPR). (Chiao 2010). Although there is no easy way to test this, the role of these genes in behavior we regard as individualistic, and their nicknames as the ADHD gene and the warrior gene might indicate that individualism was selected against in rice crop cultures not only culturally, but also biologically. Markus and Kitenawa (1991) uses the differentiating names of independent self and interdependent self to refer to that distinction, most visible in biocultural differences between East Asians and Americans. In doing so, the notion of individual itself is tensed.

In philosophy of mind, the notion of an individual is also not without its tensions. In *The Self as the Center of Narrative Gravity* for instance, Dennett conceives of self and individuality as being an *abstracta*, an abstract concept that we use to make sense of the world and facilitate calculation to better assess the functioning of others and of ourselves, not an *illata,* a thing that exists/stands by itself. This relation of an individual building itself - as opposed to being metaphysically determined by some discontinuous property in its biology, physics, or computational structure (Searle 1980) is described in detail in *Making Sense of Ourselves* (Dennett 1981) and *I Am a Strange Loop* (Hofstadter 2007). These are tensions in anthropology, biology, and philosophy, on the concept of individual itself.

Yet altruism and cooperation go, by definition, beyond the individual. This

expanding circle of entities we are willing to provide resources to was conceptualized in the human case as being a "circle of empathy," originally a philosophical notion whose representational structure we can now understand through a more neuroscientific analysis to which we now turn.

Expanding circle of cooperation, also neuropeptides

The expansion of our circle of empathy Peter Singer (1964) goes through a multi-stage process, usually some philosopher conceptualizes abstractly that maybe we should consider, say, the neighbor tribe to be human and in some sense worthy of our empathy and caring sensors. This is followed by a cultural shift where the consideration of a group as other is slowly dissolved within a culture. In the case of western Christian culture, there is an idea that even "an other" from "a *far"* (Hanson 2012) tribe has a divine individuality (Peterson Youtube biblical lectures 2017) to her. Even the criminal, the whore, the thief have a spark of the divine which makes them worthy of attention and care at some cognitive level. And yet, our amygdalas resist (Sapolsky 2017). Show someone a picture of a face from a different race, and their amygdala, a region often activated in covariation with fear and anger, automatically light up briefly, prior to being tamed - inhibited - by the manual mode (Greene 2013) coming from the frontal cortex.

Yet time moves on and the idea of a philosopher gets inscribed into culture, and then the culture reinforces and rewards those whose circle of empathy is a little bit larger, both in geographical terms - Brazilians are people too - and in biological distance terms - even people whose body has hair deserve our compassion and understanding.

At least so far in the last 4 thousand years approximately, we have been expanding more and more the magnitude of our circle of empathy in the evolving - in the Darwinian and not progressivist sense - culture that starts in ancient Egypt and became Western culture. Whereas once only the Pharaoh was inscribed with a spark of divinity (Peterson 2002), as our mythology changed, more and more fractions of the population were thought of as having an invisible property that secures them some sort of moral worthiness, either a divine light, as in Christianity, or inalienable rights, as in many modern nation states, or simply being considered "one like myself" in many subcultures of contemporary western civilization, such as the United Nations. We have assigned moral patienthood to more and more entities over time (Singer 2015).

Biology itself does respond to these changes though, specifically, the neuropeptides vasopressin and oxytocin (Bartz et al 2011a) tend to increase the acuity with which we cooperate and collaborate with those we see as "us." These cuddly cozy hormones and neuropeptides of love aren't as hippie as we may think though, as they prompt us to judge "others" more harshly than we did before, or at least discriminate more thoroughly some individuals as being "others" (Sapolsky 2017). Higher doses of

these neuropeptides and hormones, or stronger sensitivity to them, cause us to strengthen the boundary between us and them, they make the circle boundary thicker, as it were (Sapolsky 2017, Bartz et al 2011b).

Slippery slope gone too far, addicted to altruism

Is there such a thing as too much altruism? Well, maybe. There are cases where altruism and self-sacrifice become positively dangerous from a philosopher's perspective. This is what happens if, for example, genetic altruism makes individuals willing to undergo agonizing pain for the good of the genes. Even though some genes get to replicate, no minds were beneficiaries of the gene's good deeds, which makes moral philosophers sad.

The most well known example of biological altruism taken to extremes is the formation of superorganisms in ants, bees, and other insect species. The plethora of group friendly behaviors that ant colonies develop has led many to speculate about a complete shift in what unit counts as the individual in the case of ants (Hofstadter 1979) - more about this line of thinking soon. Interestingly, so called eusocial species seem to have become dependent on their altruistic eliciting genes and behaviors. It is no longer optional to them to behave in such ways and the superorganism as a whole cannot survive if its parts, that is, the individual ants, don't do their self sacrificial parts in preserving the cycles of constraints and development that constitutes a colony. Individual ants of the vast majority of species of ants have lost their autonomy (Hölldober & Wilson 2008), and are completely dependent on a network of altruists surrounding them, they are locked in.

Perhaps surprisingly, we too are locked in in this way. Deacon & Hui (2010) examine this event by drawing a comparison. At some point in the development of our species from our monkey ancestors, we lost the ability to produce vitamin C after which our rhodopsin genes - the genes that make our eyes see more colors - duplicated which allowed differentiation and eventually made us better at more accurately distinguishing between wavelengths. This is exactly the adaptation that a primate would need to find the types of fruit where vitamin C is abundant, thereby eliminating the need for endogenous production of it. The fact is that we, and independently a couple other primates in different lineages, got ourselves the ability to find vitamin C to a point where it filled the functional role previously exerted by endogenous production. We can't live without it, and we no longer make it.

Now in the case of humans and altruism, our species has been in the business of domesticating itself in many different senses (Sánchez-Villagra, M. R., & van Schaik, C. P. 2019). We give strangers information on the street, help old ladies cross the road,

donate to research in artificial intelligence to prevent the world from being destroyed by an ill designed superintelligent system - well, maybe that is just my friends - and we have progressively made ourselves more childlike, neotenous (Bromhall 2004) as well as less aggressive.

In their *The Evolution of Altruism via Social Addiction,* (2010) Hui and Deacon hypothesize that we went overboard. Now, they claim, much like many an ant, we too have let go of our autonomy and are fully dependent on each other's altruism to survive in the world. Thoreau notwithstanding, most of the interconnected people today have slid a bit from autonomous to dependant.

Yet, just like we can easily access vitamin C from fruit production and other crops, most humans have a healthy access, direct or indirect to a large amount of other humans and therefore no need to fix that particular addiction immediately. We can assume one of *us* will be there to help us, if not in times of dire need, at least in regular non dire day to day situations.

Autonomy and Individuality

The notion of autonomy can be deeply connected with individuality. Autonomy is however always partial, and much like altruism admits multiple conceptualizations. Here I'm thinking to a great extent about the way physicist Max Tegmark (unpublished) conceptualizes it, as having one's future survival and continuation being exhaustively determined by the physical state of that agent in the past. Tegmark introduced that concept in the context of artificial agents to try and find a boundary condition or something close to it to separate the current narrow AIs we have driving cars and an AI that sought its own goals independently.

From this agent based, theoretical model of autonomy we can conceptualize evolutionary or reproductive autonomy as being sufficient to cause the next cycle of one's reproduction, by which criterion ants fail even as a male-female duo, since the reproductive unit requires a colony and queen.

Humans, in creating the interdependencies that organize our social living, have lost some level of autonomy and become more altruistic and more cooperative.

Consequences that matter: from cooperation to utilitarianism

One problem with cooperating as a strategy to be altruistic is that it is nearsighted. If I cooperate with you, or with the beggar in the street, you will be better off or the beggar will, but plausibly the same resources used to cooperate in that interaction could have been more impactfully used elsewhere. Due to economic inequality, there are almost no cases where it makes sense to give beggars money you

could counterfactually give to third world poor teenagers, prima facie (Singer 2015). You can help one person eat one meal today, or you can pay for food for a family for a week far away for the same price. Most people would pick the latter in an abstract question set, and most people would pick the former in reality, which is what they do every time they give to beggars.

If what matters from an altruistic perspective are the consequences themselves, and if we are trying to maximize the value of the sum total of happiness or good accrued by everyone (global altruism)(MacAskill 2015), we can outperform ourselves by analyzing the consequences not in an interaction by interaction basis, or worse impulsively letting our non cognitive unconscious make the distinction. We can instead try to analyze the consequences directly.

Why would we do that?

Let us examine first why not let our altruism hormones decide for us. De Dreu (2011; Bartz 2011a; Bartz 2011b) has shown that Oxytocin increases parochialism, favoring one's in-group. In the paper *Oxytocin Promotes Ethnocentrism in Humans* (de Dreu et al 2011) It may play a role in detecting who is an *us* and who is a *them*. A useful mechanism in the Savannah when your tribe is running out of food but not that relevant in a globalized multi-ethnical world. Most people agree that in-group bias is bad, in particular in group racial bias - racism. There may be some value in keeping separate genetic clusters of populations, as it increases diversity and therefore resilience of the system as a whole to pathogens or some other kind of biological disruption, although it is far from clear that preserving all races in their current form is necessary. We will discuss racial bias, the good, the bad, and the ugly at length in other sections of this writing. For the time being we can keep in mind that oxytocin makes us more parochial (Fabiano, J. 2020) and, therefore, more racist/ethnocentric (de Dreu et al 2011; de Dreu 2012).

A more thorough examination of the role of empathy and compassion in the process of helping others is Keltner et al (2014) which provides a historical overview of intrapsychic, dyadic, group, and cultural components of prosocial processes.

Kicking the ladder: what got us here won't get us there

Contrast that with actually thinking about the consequences of taking action A or B. The main advantage here is that our cognitive minds can conceive of actions whose effects play out in the long term, which have ripple effects, or that can benefit parties that are not directly visible. In each of those cases, automatic mental modes cannot be trusted to make the choice that benefits more the most (Greene 2013).

Evolution gave us a ladder that correlates our destinies (Wright 2001) increasing how much organisms can benefit one another by being mutually helpful instead of harmful. The ladder of emotional empathy. The quantity of entangled organisms whose destinies have become correlated has increased almost monotonically since the beginning of pluricellular life (Smith & Szathmary 1997, Wright 2001). Now we have brains equipped with the ability to empathize (Anderson & Keltner 2002), backed by mirror neurons able to emulate, and procedural memories capable of imitating, on top of our natural nurture circuitry. One would think this is excellent news. The more the mind progressed, the larger our empathy circle became. Not so. Due to territory, food, and female access constraints, there's always been pressure to determine friend from foe, ally from enemy and *us* from *them*. We may exist in circles of sharing and cooperating, but much like the outmoded model of atomic layers, we have strict tiers and have a proclivity to react very differently to kin, probable kin, surrogate kin, friends, acquaintances, neighbours, others and outsiders. In fact the whole notion of a *circle of empathy* (Singer 2011) should be substituted by *circle of paranoid super precise alliance formation cognitive schemata* (DeScioli & Kurzban 2013). Our cognitive automatic moral mechanisms are much closer to FBI agents than to hippies.

Roughly speaking, evolution gave us a ladder so we can cooperate, but it is time to kick the ladder and go straight for the prize, to cooperate not in alignment with our cognition, but *despite* our moral circuitry. The main reason for this is the incredible expansion in the size of our world and those we consider morally worthy of our attention. Not only brothers and sisters, not only people in the present, not only people alive today, not only people in our race. For many, not only people, but also animals. For some, even artificial intelligences (Bostrom 2013). Going even further along this spectrum we meet those who think all reinforcement learning systems have something worth protecting, and the very edge, the pan-psychists, who think electrons are people too. Well, at least in the sense of moral patienthood. For fascinating discussions on these topics it is worth consulting the essays on reducing suffering, work by Brian Tomasik (articles at reducing-suffering.org) on how far we should extend our deliberate cognitive empathy circuits as we realize how flawed the automatic ones are.

This "let us kick the ladder" point is important enough that I want to give a few examples of it. For one thing, Harvard philosophy and neuroscientist Joshua Greene dedicated a whole book, *Moral Tribes* (2014) to the issue. In the book he calls one of these modes "manual" and the other "automatic." The classic example of letting go of our automatic mode is pushing the proverbial fat man off a bridge to save five people in the Trolley problem. Let us think of some more probable examples in the day-to-day life of a normal human, for the sake of usefulness.

Obama's Suits: In our opulent and affluent society most people's main constraint is quantity of attention, and amount of brain power allocated to decisions, which can be

limited by not allowing one's future self to decide. Barack Obama for instance only had two colors of suits so he wouldn't have to waste his valuable decision power in picking up clothes in the morning. Similarly the mere act of not using Facebook, or not entering a shop where there's many products available can end up being an altruistic act for someone who spends another part of their time allocating attention to making decisions about the most effective altruistic causes they could support and how to go about doing it.

Career Choice: Most people would also not consider choosing a career a potentially altruistic act, but nothing could be further from the truth, there are two very substantial aspects of a career that may account for the majority of the altruistic act when executed over a lifetime:

1. externalities, and
2. potential for future donations.

Value Capture: In the world of entrepreneurs two things are key for those who want to be financially successful:

- to produce a large amount of value, and
- to be able to capture that value,

Not so for altruists, who are not deterred by the second constraint. Since the competition for creating value that can be captured is fairly cutthroat this gives altruistic minded people a massive advantage in that they can just create mechanisms and systems where it's impossible to capture the value but there are enormous gains to be had which are distributed over a large population in small quantities or even a smaller population but in large quantities. We can call this the "externality leverage". The most obvious case of an externality leverage is the benefit accrued from scientific discovery that can prevent disease, it's usually a very hard value to capture but it provides unbounded potential of happiness and well-being to others (MacAskill 2015).

There are plenty of real life examples one is more likely to encounter than obesity adjacent to trolleys and kidnapees tied to train tracks, and in those cases it is important to not let the automatic mode (Greene 2013) decide if one wants to make the truly altruistic decision.

There is however a different kind of automatic that will inevitably become a larger and larger fraction of interactions and of the economy in the coming decades, and has for 80 years already. Artificial automata, colloquially known as computers.

The Golden Rule in Virtual Worlds: good and real philosophical robots

Game theoretic robots and Tit for Tat evolution

To many, altruism isn't properly human in as much as humane, which brings us to the question of what kinds of other entities can be humane without necessarily being human. Although animals, including our closest relatives (Woods 2007) have been shown to be altruistic in some situations (De Waal et al 1973), we will spend more time considering a different kind of system and its ability to give and take: computer programs.

We will consider three kinds of computer programs, virtual robots, robots, and superintelligences, to be defined later.

Virtual robots have been used to assess strategic outcomes in game theory for many decades. In the most common usage, researchers are interested in verifying for a pair of robots, what would be the outcome of a dyadic iterated interaction between these virtual bots. That is, if you made them play a game where they can e.g. cooperate, defect, and see what the other did, and experience the outcome of previous rounds, which strategies are more successful?

Success is often conceptualized with a cardinal payoff matrix, or less often an ordinal one, where the outcomes of all combinations of action are displayed.

Two robots that cooperate on every round (a human observer might consider calling them "utopian optimists") would get the best possible outcome, but if one of them defects on the last round (gotcha optimist), he could get an even better outcome. But then the other one has an incentive to do the same (becoming a suspicious optimist), which brings the incentive backwards one round, then the other has the same incentives, all the way back to round one, in which case maybe just defect on every round anyways (nihilist cynic).

But clearly there's a problem with being a nihilist cynic robot, especially if sometimes you are playing against robots that are not defecting all the time. So what if you are in a competition and you have to average the best result from multiple sets of 100-round interactions with robots that are not you - well, we can throw robots that do the same as you into the mix as well. What happens then? This competition became famous decades ago when it was found that the strategic robot TFT, tit for tat, was the winner of one of these artificial competitions. TFT cooperated on the first round, and then did whatever their partner did from then on - I resist calling the other participant

"opponent" because two robots gladly cooperating from round 1 to round 100 don't strike me exactly as an opposition, and the game structure is clearly not zero sum, where for one to win, another had to lose an equal and inverse quantity.

TFT won, making its strategy famous. A little too famous because that information trumped much of the later developments in the field, which expanded in three main directions:

- It complexified the situation, with robots allowed to do things like randomize aspects of their strategy, signal honestly or dishonestly what they were going to do, and more recently even scrutinize the source code of their partner directly (Taylor et al 2016) that is, fully read the program which the other is instantiating.
- It complexified the number of agents in each interaction and allowed for punishment of others for a fee and other scenarios that more closely resemble the real world, where we are preceded not by 99 other necessary interactions but by our reputation, and whatever action commitments it entails in whoever we choose to partner with.

What has been learned from these complexifications is that there are different strategies that can be added to assist cooperation when you change the game.

If you change the game for information to be mixed up with noise, that is, sometimes you perceive the other agent as cooperating when they didn't, and vice versa, then the strategy that counters that is *forgiveness.* Forgiveness is cooperating with some probability even in the face of (perceived) defection (McElreath & Boyd 2007). The robot equivalent of showing the other cheek perhaps.

If you change the game to have freedom of association, and allow for some level of reputation transmission, then agents who freely choose to associate with other agents that are cooperative, to the exclusion of defectors, manage to both stabilize their groups and gain benefits of the game they are playing if it has a bounty or benefit. This has recently been shown to be why human neighborhoods also tend to cluster around higher and lower levels of trust, with the lower trust ones becoming more problematic over time and the higher trust ones more desirable - this effect doesn't have to be only cultural or individual choice, genetic assortment also influences personality characteristics which might cluster high conscientiousness people together, leading to a more stable environment in which to trust, which leads to more trust. Some even argue that at a much larger geographical and temporal scale, the economic meltdown of Greece 2007 and its rescue by Germany is similarly associated with this high trust low trust contrast.

If you change the game to allow for dishonest signaling (deliberately claiming you cooperated even though you defected in a way that is time delayed, similar to what the Ponzi scheme agent does) then suspicious strategies may evolve, or even paranoid

ones. A suspicious agent may start by defecting for some number of rounds. A paranoid agent will occasionally leave a group towards another one even absent signs of defection, or will simply act in the way it does when other agents defected sometimes even if they didn't.

The labels I'm using here are descriptive of the labels we use to designate other people when they use similar strategies to those used by the computer, though there is no implication that the computer or the computer program would have an experience of paranoia or suspicion, that is for philosophy of mind to decide, and that is a field I have left because it does not seem it will anytime soon.

Trustworthiness in humans and machines

Different from the simple machines that execute TFT and related strategies, humans exercise trust heuristically. There are too many available sources of information to determine whether someone is trustworthy, all of them probabilistic in nature, and thus, we need to combine them to make an assessment and that assessment may turn out wrong even if we calculated the probabilities correctly - in the same way some points in a Gaussian fall outside 2 standard deviations from the norm. If all you know about someone is that you first saw them running away from two cops while a very loud noise came from half a block away at the federal maximum security prison, and the two cops were shot down by snipers, at which point the person started walking slowly before talking to you, you would be less likely to abide by their request "Do you mind asking for an Uber for me please?" than if they were say the mayor of your city leaving town hall with a preoccupied semblance.

Some heuristics we use to judge a person's trustworthiness are considered undesirable, even sometimes illegal. Genetic proximity, favoring one's family members, or disfavouring certain races (technically clades) is frequently illegal or frowned upon in western countries. Political nepotism for close family members is illegal in Brazil, and many countries have laws of illegal succession from father to child of elected cargoes. Discrimination in virtue of race is illegal in most of Europe, Canada, and American states. Other heuristics might come from smell, manner of walking, table manners, diet, dress code etc... and being mutual acquaintances of a trusted third party. These are usually not illegal and might be too subtle to be easily detectable and frowned upon.

We don't only track trustworthiness by assessing the reference classes to which an individual belongs, we might also obtain information about their past history, from curriculum descriptions to previous landlord references to criminal records to veteran cards. Those pieces of information directly about the individual provide us with additional information to what we initially sense with automatic heuristics. Third party

information completes the constitution of what we think of someone and informs our choice of cooperating or not with that person.

Trust is largely less understood than other game theoretical strategies and schemata used by humans - though see Zak (2017) for a neuroscientific perspective. Unlike personality and intelligence, where we know hundreds of the specific genes, the loading of genes and countless other correlates to the point where we can now gage the genetic component of intelligence of any ancient population from a set of genomes alone (Woodley of Menie & Figueiredo 2013), we don't understand very well the correlates of high trust and lower trust societies in humans. In part for this reason, trust isn't one of the categories used to program virtual agents in game theoretic competitions and simulations.

Treating others as you'd like to be treated

A free translation of the Sermon of the Mount would read: Cooperate with others unless you are costly punishing them for behavioral reasons in a way that you would deem acceptable being punished yourself to enforce individual, contractual or moral norms that promote long term satisfaction to the parties or the community.

To say "one should do unto others as one would like them to do unto you" has a myriad of problems to which we can attend now:

- "One should do unto others as one would like them to do unto you" doesn't work because *people are different*
    - In particular the two sexes have been doing what is biologically describable as a cold war for 500 million years for returns on parental investment (Trivers 1973, 1976; Buss 2005) and thus our psychologies have been optimized by evolution to be at the same time complimentary for some tasks (e.g. navigation based off reference points plus navigation via geometry beats either one alone), and competitive in others, such as courtship and mating which are for the most part positional goods even in societies where people have multiple mates, as both male attention and resources, and female pregnancy, are scarce resources.
    - Another general class of difference is neurodiversity, which makes some people thrive in environments that others may find abominable, and vice versa. Psychological conditions such as Asperger, ADHD, bipolar, OCD and others make it the case that to treat others in a way that best harmonizes with their emotional systems and cognitive capabilities requires treating them differently from oneself.

- "One should do unto others as one would like them to do unto you" doesn't work because *counterfactual actions matter*
    - The moral value of one's action isn't only predicated on the amount of good it generates further downstream a causal chain, as if someone else would have taken the same action if you didn't, then the causal chain would mostly be preserved. The moral value of one's action is better thought of as difference, the difference between what would have happened if one took or didn't take such a course of action. Preventing the spread of CoVid19 through washing our hands and general societal panic reduces contagion at an early stage, which later prompts some to call those who worried early paranoid if containment is successful, yet it may be the very action deemed paranoid that prevented the tail risk scenario of a second Spanish Flu, only deadlier. (I write this 13th of March 2020)
- "One should do unto others as one would like them to do unto you" doesn't work because *anomalous causation is poorly understood*[9]
    - Scenario: you know ten copies of you will be created during your sleep, and you'll be awake the next morning in rooms that look the same, except labeled 1 to 10 on the outside, where you can't see. Upon waking up you should assign a 40% probability to being in a room whose number is smaller than 5.
        - What if you don't know how many copies were created, but you do know they have numbers, and you wake up and find yourself in number 100? You can be sure that the total number created was not 99, and you'll have a decreasing probability from each number starting at 100.
        - So if you are the 100 billionth mind with symbolic capability (Deacon 1997) to live, that is, if you are the 100 billionth member of the symbolic species, how many members of the symbolic species should you infer will exist overall? Much like the example in the room, a decreasing probability starting with the historical total until this point. Modulated by your knowledge about the contemporary world of course.
        - If meteors were to hit the earth with 50% probability according to the estimates of physicists, we could try to game the probabilities using the ideas above (Bostrom 2001). To do so an international organ such as the United Nations could unanimously vote that if the meteors hit, the human species would resolutely decide to multiply

[9] This is not the only way of dealing with problems involving self selection and anthropic reasoning, for more complete accounts see (Grace 2010, Bostrom & Circovic 2003, and Olum 2002)

itself 100x and expand into the universe and will do nothing else, full throttle for the next two centuries. Since the probability that the total number of symbolic specimens to exist being 100x more is 100 times less probable by the reasoning above, this would potentially give us reason to expect the meteors to hit the earth with much lower probability. That is, having a concept of "timeless probability" over "total number of symbolic specimens" allows us to game physics by making resolute decisions to create many more of us, which is a very improbable scenario given our position in the temporal order.

- It is not clear whether probabilities can be conceived of in this timeless manner (Yudkowsky 2011). But the arguments against that type of reasoning usually have flaws that run as deep as the apparent flaws that this type of reasoning seems to have. It should be expected of course that arguments that do not involve survival in the Savannah, or the farm, or even in an industrial society would sound counterintuitive to 21st century ape brains, so the mere accusation of strangeness does not suffice for us to eliminate anomalous causation as a potential source of moral action. We also refrain from blaming the biblical prophets for not adding this corollary in the original Sermon of the Mount, likely for mnemonic reasons.

- Reference class tennis
  - The problem with arguments like the above, as well as simpler questions about determining one's individual chances of having cancer based on statistics of different groups one belongs to, is that no one knows how to choose which reference class to pick from. As a 31 year old male, Should I assess my likelihood of heart attack from the 30-40 male group, or from the 25-35 male and female group? What about the study which analyzed it based on height, regardless of age and sex? And if I am to convolve these distributions, which sort of convolution (the operation of joining distributions) operation to use, to assess my own probability?
  - These questions may seem initially abstruse and unrelated to altruism and cooperation, but in many ways they are not. As we have seen in other domains, determining if someone is one of "us" or an "other", for the purposes of a particular interaction, seems to be one of the chief characteristics that establish how humans interact with one another. Oxytocin modulation seems to help us

sharpen that distinction, making it even more salient (Bartz et al 2011a)(Sapolsky 2017). The response to that question is usually the difference between cooperation and withdrawal or defection, and stable groups often depend on it for their survival.

- I have used *symbolic specimen* as the appropriate referent for questions of anomalous causation and reference class, because the symbolic capability seems to be very discontinuous (Deacon 1997) separating humans from any other living animals, and the ability to manipulate and represent symbolically - or to emulate such representation - seems to be what enables the level of technological prowess that most benefited our species so far, as well as the ability that could lead us to prevent existential risks and other extinction events, making us (the symbolic species) the only known beings in the universe with the ability of creating astronomical flourishing (Bostrom 2003) for humans, animals, and - if they too are moral patients - machines as well. Ought implies can, and if any species ought to be morally responsible for the destiny of living beings on earth, it ought to be the symbolic species, the only one that can. This puts us at a morally higher and distinct level above all other contemporary life and machine forms in our moral responsibility.
- The other problem that intersects and complexifies the situation even more here is that it is not clear what temporal scale we should use for arguments that rely on a reference class. Above when saying that you are the 100th billionth symbolic specimen around I was arbitrarily assuming personal identity over the course of a lifetime, instead of, for instance, each time you wake up until you sleep again or other ontologies.
- Personal identity seems to be a question that breaks down into three other ones, according to Parfit's *Is Personal Identity What Matters?* (2011):
    - Psychological continuity
    - Memory continuity
    - Causal continuity

  Once you know the degree of these three metrics, there is no further question to be asked. "But is it me?" adds no additional information to the three levels of continuity above.
- The problem then is that this still doesn't tell us the degree of resolution of the time slices you are comparing (Lewis 2002), and the same problem happens in questions about picking a reference

class to reason about anthropic arguments, such as those about anomalous causation above. Questions such as these:

- I am only sure that I exist now, should I not consider just a three seconds slice of me as a unit for reference class forecast, in which case I'm not the 100 billionth, but plausibly the 100 quadrillionth 3 second slice of symbolic specimen there is?
- The shorter the timespan of what I consider me, the more weird scenarios I can conceive of for the experience I'm having (brains being spit from black holes for instance). Cartesian certainty without a hefty dose of Humean induction perishes into Wittgensteinian subjectless-tivity and on and on all the way to Boltzmann brains.

The overarching theme that unites the bullet points above is that individuality and altruism are deeply interwoven with complex notions of self and other in ways that, simply put, cannot be compartmentalized and cast away from careful examination. The idea of treating others as one would like, be it in the religious framing, the golden rule, or other formulations, are all weak approximations, or coarse grained description of phenomena that admit of multiple compatible models (for those familiar with cardinality mathematics, this problem is analogous do Löwenheim-Skolem and infinitarian models).

Tying this to our principal question of scalability, different approximations of what a self or person is will render different possible methods and strategies for larger scale altruism, in both time and space. We need a deepened understanding of the nature of these questions in the intersection of philosophy of mind and game theory of agents, if we are to have clarity about how to steer a more altruistic whole.

Good and Real philosophical Robots

The field of agent interaction in strategic games with different constraints is constantly evolving, and becoming more connected both with the considerations coming from the biology of cooperation and altruism, Bowles (Bowles & Gintis 2011) has done interesting work in that intersection, and more recently there is also work connecting more and more to the field of Artificial General Intelligence, where arguably it moves from philosophical curiosity to a life or death subject.

Tit for Tat that reads code

One domain that has been developed more recently is the strategy development for agents in a competitive environment that can read each other's code. So it can

basically calculate what the other agent will do if they cooperate, defect, leave, etc… And the question of how to design effective agents in those circumstances ties in to many possibly relevant considerations, of interactions between machines and also between machines and humans

Lie detectors

In some domains, ad targeting algorithms already know us better than we know ourselves (Harari 2021. Ward et al 2017). Many women found out they were pregnant because the algorithm of a social media website started showing them baby related content. As this process becomes more and more sophisticated we are reminded of lie detectors, which still are not perfect but perform strongly above randomness. In a program that reads another's code, that transparency would prevent lying altogether. The absence of lies reduces the cost of signal errors, by making moot the question "did they really defect, or did I just detect a defection when a cooperation happened?" all you have to do is investigate (read the code, or ask or check the detector).

Superrationality and Timeless Decision Theory

Hofstadter (1985) suggests that a principle of superrationality is in order, whereby we assume another will make the same decision we will. He aims to circumvent the problem of defection in the prisoner dilemma, he wants to argue that cooperating is the rational thing to do. Or at least the superrational thing to do.

While we can be agnostic over the metaphysical strength of his argument, we can note that the more we think an agent is us, the more likely we are to behave superrationaly in relation to it. We cooperate with the person in the mirror. We cooperate with our future selves, and in abstract scenarios we tend to favor ourselves approximately the same amount.

If Hofstadter can extend the principle of rationality to superationality, Yudkwosky attempted a similar move (2011) with *Timeless Decision Theory*, developing an entire decision theory in 100 pages which doesn't have time related inconsistencies. The algorithm decides what the agent does in a timeless way. Again this depends on believing that in all ways relevant for decision making, the agent making the decision at all times is the same agent.

Yet the electrons moving in a computer making the same decision in the same program via the same algorithms are not the same electrons. So, is it the same agent in the relevant sense? How long a time-step sequence needs to be to determine that two algorithms that react alike are the same algorithm?

These questions substantially complicate the questions of altruism in the case of humans. But we can say initially that there is some reference class narrow enough given

an agent such that this agent can be considered the same as another. And when that happens, we should desire that the agent act superrationaly and timelessly.

Levels of sameness and detectability of sameness.

Being the same as something is a complex philosophical notion. To bypass the philosophy we can say that for all practical ethical purposes, an agent is the same as another if all agents that are partaking in transactions or interactions with it cannot distinguish them in a behavior influencing way, including themselves. That is a verificationalist notion of sameness.

Acausal trade

A more exotic consideration when it comes to altruism and cooperation is the idea of acausal trade. Supposing I value oranges existing but can make cheap apples, you value apples but can make cheap oranges, and we are causally disconnected from each other, but we know of the other's existence, we could infer that we are both better off by making what the other values, similar to the prison dilemma. Acausal trade is particularly prone to predation by the *Pascal Mugging* argument (Bostrom 2009), i.e. arguments where an infinity or Very Large Number is supposed somewhere and you are persuaded to behave in such way as to attain the infinite value with some probability, but then risk losing all the finite guaranteed value you could obtain. We will return to this in the section on infinity shades.

The takeaway from this exploration into altruism in AIs and artificial agents for our original question of scalability is that even though mechanization offers a promising solution for creating equilibria in which altruism is facilitated, as is cooperation, many technical specific questions remain, the hardest of which to solve being the question of identity in artificial agents, which ultimately determines if an act is egotistical or altruistic.

For now we let go of the domain of artificial agents and artificial intelligences to delve back into how we think of altruism. What makes us even able to be altruistic psychologically and cognitively, and what do we mean when we act in such ways.

Metaphors and Analogies: asymmetric conventions and the shape of modal thought

Meaning by similarity

One way to understand the notion of what something means is by similarity (Hofstadter 2007). Whenever we say that chocolate mousse is like chocolate tart, except it is just the filling without the crust, we are using our ability to abstract some properties and features out of a specific entity or situation, and assume that some of those properties are present in this other situation or entity. Much of our understanding of the sciences is couched in similar processes (Lakoff 1980). There are several reasons why we use this shortcut (ding!) to explain, to think and to reason. Our minds are at their best when dealing with mesoscopic entities in the day to day world, preferably in situations that interfere with fitness, such as social situations. That is, we are most at home (ding!) when dealing with the kind of problem that our brains and those of our ancestors have been solving for aeons, problems that our sensory apparatus is equipped (ding!) to tackle, and that we have enough of a grasp of (ding!) to enter head-on (ding!).

Individual meaning by similarity between minds

A process like translation happens not only between texts in different languages, but also in processing the same word by different minds. Though our conceptual networks are somewhat similar when we are speakers of the same language, there are always nuances and particularities in how we take secondary connotations of words, sentences and phrases. The process of transforming a string of symbols into a tree structure with grammar that lights up (ding!) a path (ding!) in the conceptual network of the information recipient (Pinker 2014) is executed through a laborious cognitive process of interpreting metaphors (Lakoff 1980) and drawing internal analogies (Hofstadter & Sander 2013).

Why does this matter to altruism and cooperation? Mostly it matters because utilitarians and effective altruists conceive of altruism in terms of counterfactuals (Lewis 1967) and counterfactual theory has been to a great extent conceived of by using metaphors of proximity between possible worlds, and using that to determine "what would have happened if you didn't do X" and then subtracting the value in one case for the value in the other case to obtain how much value was created by your action (Bostrom 2011, *Astronomical Waste*).

So when we ask what does it mean "you saved 4 lives" if everyone dies one day, a plausible philosophically accurate response is: I have changed the world in a such a way that, the average value of number of people alive in the set of worlds that I cannot epistemically eliminate as the world I'm in, subtracted by the average value of of number of people alive in the worlds tied for closest possible world where I did not take that action, equals 4 people, where "people" stands for something like 40-100 years of the relevant kind of moral patienthood and moral agenthood that would qualify as a person under normal conditions.

The notion of counterfactuals is very present in the Effective Altruism *ethos* as well as in the *ethos* of utilitarian philosophy. For those who think that consequences matter, it seems to frequently follow that the consequences that would not have happened otherwise matter more, and frequently are conceived of as the *only* consequences that matter, leaving causal overdetermination out of the picture. Alternatively, one could say the consequences that matter are those that are *expected at the time of action decision*, with some probability, to not be over-determined. So if both of us give the same child a virtual copy of the book *There's no Such Thing as a Dragon*, these actions would count as altruistic if and only if we didn't know the other one was doing the same thing, or if our actions were otherwise causally entangled, such that each of us could only do it if the other did.

Since altruism is conceived of frequently in terms of counterfactuals, and counterfactuals are conceived of in terms of possible worlds, and possible worlds rely on cognitive metaphors and how they are executed in our minds, we will now turn our attention to the nature of cognitive metaphors, primitive metaphors and how cognition utilizes them, and to what extent we use them to assess actions as altruistic, and to judge others morally for the actions they execute.

Modal thinking operates by intertwining cognitive complex notions of proximity with frame-based metaphor-like notions of proximity.

Minds as Metaphorical Engines: embodying cognition

Are we doing altruism right?

Effective altruists and utilitarian thinkers attempt to calculate the differences between worlds in which they took or didn't take an action to figure out the moral value of that action. But what are these possible worlds we conceive of where we didn't take actions we in fact took? Or where we took actions we in fact didn't?

There are two questions that can be asked about these worlds: how do we conceive of them, and how does that affect our altruistic actions?

And: what are they really?

The nature of possible worlds is an undecided topic in metaphysics. If they are in the same sense our world is, a proposal famously endorsed by philosopher David Lewis (e.g. 1986) throughout his career, then which action you take doesn't change which worlds that are. The set of all worlds that are is fixed, so the amount of good or bad there will be in the whole of history across the multiverse of *possibilia* is also fixed. If there are infinitely many worlds, then there will be infinite good and bad, and finite actions do not change the total sum of good and bad. For a more extensive discussion of apparent

paradoxes of infinitarian ethics and some proposed solutions involving hyperreal numbers, see Bostrom (2011).

Whether or not possible worlds exist it is a fact that we can think of worlds different from our own such as Sherlock Holmes's world, although when we do so, we do not think of those worlds exhaustively. That is we do not conceive of them with all their nitty gritty particularities, but only that which is relevant to understand that fiction, that hypothesis, that alternative, that story. Linguists note that often we conceive of conceptual blends (Fauconnier & Turner 2008, Hofstadter 2013) which only use a few objects to imagine a situation being altered. In theory, altruistic agents would want to calculate at least approximately the whole causal structure entangled with an action. In practice, this is too hard, as ironized by Jordan Peterson in *Maps of Meaning* (1997), so we just estimate based on what we know. Some abandon this action based paradigm altogether and instead go with rule utilitarianism. That is, instead of trying to compute or calculate at every action decision node that seems relevant which action to take, they try to come up with a rule system that, as a whole, seems to cause the most amount of good if you follow it. This reduces the time, attention and computational cost of having to find out what to do every time, substituting it for finding out which rule to implement every time, from a set that you created when you had a little more time to reflect on a good system of rules.

Cognitive Metaphors

The cognitive metaphor system we use to think linguistic thoughts can be decomposed into subtypes:

- Primitive metaphors:
  - Cognition is embodied in primary metaphors that primarily specialize relations between things, and put them in and out of containers. These are learned when we are relatively young and imprinted in our cognitive architecture, they are a structural scaffold of metaphors through which we build many other metaphors.
- Asymmetric analogies
  - That is the umbrella category involving both primitive and learned metaphors. Whereas an analogy can have two items that are symmetric, a metaphor has a source domain and a target domain, and the asymmetry is constitutive of the metaphor. A is like B in some way, and B is something we are already familiar with, and we know which are the salient features that make it like other things. We learned it.

- Cognitive Metaphors related to moral reasoning
  - Narrowing the vast domain of metaphor, we have cognitive metaphors related to moral reasoning, which are those we use in order to conceive of something as moral or immoral, ethical or unethical, cooperative or conflictive, altruistic or selfish.
- Frames
  - A Frame (Fauconnier 2008) is a conceptual entity that is presupposed by a narrative or a metaphor. Instead of floating in the void, a metaphor tends to create a conceptual assumption of a domain in which to think of it. This could be a scenario or an environment, or an abstract structure in which the metaphor makes sense. It's the set of underlying assumptions and their consequences to make the metaphor meaningful.

Moral Metaphors: From Families to Fables

If we want to think about altruism and cooperation, moral metaphors will be particularly handy for us to understand how we think about these concepts.

- Moral credit

We often think of morals as a quantity that we have in a container. If you are more moral, you add more to your moral bank account, you get credit.

- Moral debit

Then if you screw up or do something bad, your brain thinks you've drawn from the moral account, which may mean you are a less moral person, or that you've gone all the way into moral debt. Which conveniently you may repay, either by taking morally desirable actions or, depending on your personal convictions, by negotiating with the appropriate deities.

- Nurturing family

Lakoff points out two fundamental structures organizing primitive metaphors which create two different lenses through which people view reality, the nurturing family and the authoritarian family. If you conceive of family as a source of nourishment, you are more likely for instance to end up in the liberal side of the spectrum.

- Authoritarian family

Conversely, if you conceive of family as more authoritarian, you are likely to end up in the more conservative anti-liberal side of the political spectrum.

- Care, Harm and Father Presence

We can tie that, correlationally, to father presence in a family. Since women score higher on the care/harm axis of the psychological spectrum by 61.5 to 38.5 in a percentile scale (Weisberg et al 2011), and males are more authoritarian, there is incentive, politically, for the left side to increase number of families where parental care is done exclusively by the mother, as the number of people who will have a nurturing frame through which to view the family will increase. The so-called war on boys, to the extent that is taking place, is partially incentivized by this.

- Justice, fairness

Our notion of justice and fairness is also based on metaphorical cognition (Lakoff 1980). If both of us have a container and there's some liquid to be distributed, we might judge whether we have a just or fair division based on whether we get the same amount of liquid.

- Notions of justice will depend on where you cut off the containers

However, our usage of these container metaphors doesn't decide what is just or fair because we may be framing our metaphors using different container owners. So if I am thinking of the containers as individual containers, and you as family containers, I'll be glad if we get the same amount, and you will be glad if our families get the same amount, even if I have more kids than you.

Part of the contemporary political battle the world across is to change the frame of those containers for our conception of what is fair and just. The social justice movement, influenced by Derrida, postmodernism and identity politics, is, unconsciously or consciously, trying to frame the containers at a racial level. They want each race to get their fair share of a resource that can be distributed. An individualist, such as a libertarian, might resist this racial framing, and attempt to bring the notion of justice and fairness back to the individual, where it lies in the US constitution. A nationalist could on the other hand try to bring it to the level of nation. There is a constant evolutionary struggle between higher and lower levels of evolving entities to capture the notion of fairness and justice, and therefore for entities cutoff at that level to receive more resources than they otherwise would.

This of course doesn't require, though benefits from, conscious agents attempting to explicitly do that. The rise of identity politics and racial politics is a recent victory of the level of resolution "race" to capture the resources that are allocated as fairness or justice taxation at any given time. To the extent that they are both fighting for the level of resolution to be the determinant, Nazis, white nationalists, black nationalists, social justice advocates, black lives matter, are all fighting in the same team. Against for instance the individual and the family at lower levels, and against the nation, continent, or globe at higher levels.

When we reason about morality and altruism, we are guided by primitive metaphors, cognitive metaphors, and these imagined counterfactuals, where we compare, according to a metaphorical standard, whether there would be "more" or "less" of the ethical thing, depending on which actions are taken.

Morals of Fables, oral tradition and memetics

Fables and oral tradition used to keep information that is relevant for some of these moral frames stable vertically across generations. When you tell the little piggies story to your kid, you are providing them with many frames and lenses through which to see how to act in the world. It allows for moral coordination within a community around some common ideas and frames. More recently after the invention of writing, and Disney movies, horizontal transmission became more common and mass coordination of morals also more common, from the cultural revolution in China to the attempt to promote feminism in *Frozen*, or, long before that, to save a sleeping beauty through a non consensual kiss.

Some of these fables, tales and stories feel to a large section of the population to "strike a chord" and affect us deeply in a moral manner that is hard to explain. *The Lion King*, the most sold movie of all time, is a famous example. Extrapolating from Jordan Peterson and based on the work of Jung, I hypothesize that these stories and tales are such that the moral frames and ways of thinking they provide us with serve some evolutionary purpose and thus have been selected. The Jungian archetypes tend to be frequent in different populations that survived the test of time, and some of those end up constitutive of a culture or a population.

Bret Weinstein (Weinstein, Peterson & Rogan 2017) points out that despite there being evolutionary incentives at play both in the construction of these archetypes present in stories and fables, it does not imply that there is moral good in them, only that they have survived the test of time. We can assume that some of these ideas are functional and morally desirable, some were desirable in a world that changed slowly but no longer are, and some survived merely because they hijacked some susceptible parts of our cognition for their own good.

The longer a fable has survived, and the more modifications it had before reaching its current form, the more we should expect it's morals and connotation to be aligned with an evolutionary long term goal. (Zipes, J. (2013). *Why fairy tales stick: The evolution and relevance of a genre*.)

The more people are interested in a particular story (*Lion King, 50 shades of Grey, Harry Potter, Lord of the Rings*) the more we can expect that it captures some aspects of those archetypes that are already entrenched in our cognition.

The Torah and the Bible have famously survived a long time and been edited

many times, so we should expect that they serve or served an evolutionary function, which might coincide with a moral good.

The Lion King and *50 Shades of Gray* may also contain these moral truths, but they can also be superstimuli that overpush some emotional or mental button in our minds.

Patrilocal primates in a species where some males go out and about to other groups, fight a social hierarchy and the intemperances of nature, and are sometimes rewarded with a sexually available female, as well as being able to go back into their own group to make trade or spread ideas could, plausibly, evolve a hero archetype. They would be particularly struck by stories of males that use their high testosterone and revolutionary desires to dabble into the unknown world of mother nature, search for treasures and visit lost tribes, and even slain the archetypal mixture of all the primate dangers (Dragons are the archetype of all dangerous-to-primate things except women. They are a reptilian snake like predator, which has legs like a ground predator, and wings like a flying predator, and breathe the natural element of danger, fire (Isbell 2006)).

So we listen attentively to hero stories about fighting dragons to get princesses starting off as a low status male. From Italian plumbers to feline princes to dwarfs (*The Hobbit*), we are fascinated by the hero narrative. It very likely has evolutionary value.

The Evolution of Religious Archetypes and Narratives

As the biocultural process evolves and unfolds, some structures become more template-like, and other structures become more plastic and flexible. While our specific utterances on a daily basis are flexible, their grammatical structure is fairly stable. In the last two decades, prof Jordan Peterson has developed an elaborate theory of human personality that organizes itself around this axis. Translating his theory into my vocabulary: Compression has been a problem in the transmission of behaviorally important information. Thus evolved children fables and adult archetypes, with complex morals and lessons about how to live. These lessons are organized such that they implement a template set of behaviors and algorithms of decision in primate brains while possessing low Kolmogorov complexity. They are transmitted optimizing for behavioral learning fidelity as well as replicative power.

What ensues is the composition part by part, the grafting and editing of myth, as well as the formation of psychological archetypes, a notion developed by Jung and Joseph Campbell in detail. Archetypes are a solution to a compression problem, much like song rhymes. Differently from rhymes though they also have another evolutionary

desiderata to fulfill, namely that of "being a set of qualities such that, when implemented in a Homo Sapiens brain, generates the conditions of reproduction of that human, their society, and that set of qualities." The archetypes that organize the contemporary West are direct cultural descendants of Egyptian archetypes modified through time. The beacon of civilization has been passed on, grafted on, and evolved in many different ways during historical and biogeographical transitions to different environments. Jung's collective unconscious (Jung 2014) is in my model the set of hardware and social modifications that makes some human groups particularly tenable to learn some versions of those narratives. *The Hero with a Thousand Face*s (Campbell 2008) is the invariant hero story, versions of which are so deeply ingrained in our brain or genome or human universals (Donald Brown 1991) (I don't think we know which yet) that they are re-emerging patterns. Convergent evolution.

Like a blacksmith slowly forging the iron until it is aligned and sharp, stories with the property of being memorable enough to linger on, and morally valuable enough to stabilize a society were carried on through centuries and millennia - some of which are now being studied by the discipline of Human Behavioral Ecology. Not only the stories adapted themselves to the evolutionary needs of groups and to the bodies and brains of their hosts, but also the minds of the hosts adapted to the stories. One of the starkest examples of that is *Stalking the Wild Taboo* (Edward Miller 1994) which goes into detail about mating pattern (personality) differences and their correlation with a myriad of phenotypic traits.

These foundational archetypes provide the implicit rules that govern the functioning of particular individual roles within a particular society, and the narratives from which they are drawn sometimes contain the explicit ones as well (e.g. The Ten Commandments, Sharia Law etc...). They contain narratives that our specific type of primate is likely to encounter: for instance the hero narrative is a metaphorical recount of the fate of a male primate who decides to mate exogamously in other group, tackles the forces of nature and slays the dragon to either get a princess or riches, which can later be converted to princesses. Males in matrilocal societies have those problems, with dragons being the reification of terrestrial, arboreal, and flying predators as well as fire in one being. Besides matrilocal primates, low status males in a flexible hierarchy might also go through the same adventure through being social outcasts: plumber Mario slays Dragon Bowser to get Princess Peach. Simba is cast into the forest to live with low status Timon and Pumbaa and has to fight nature and Scar to retake the kingdom and get princess Nala. Variations on these stories are rediscovered all the time, which is why Peterson, Jung, Campbell and others assume they are archetypal in a profound sense. I dispute the hypothesis that the collective unconscious is the same for all humanity though, as the evolutionary divergence of different groups into different locales and different ecologies, food density, population density and heat levels probably branched

off selective forces making archetypes differently salient and possibly differently structured for different people.

Especially after the second world war, mythological and archetypal narratives captivated young individuals (*Lord of the Rings*, *Narnia*, Disney cartoons, superhero comics, Marvel movies). Mythological group affiliation nowadays is considered a type of subculture affiliation, not a full fledged national identity - the odd *Star Wars* obsessive fan notwithstanding.

There's a complex chicken and egg problem in determining if an archetypal story sells more because it is archetypal (say, Disney's 1994 *Lion King*) or whether we are tempted to look for archetypal foundations in best selling stories or long lived stories. Regardless, the idea of archetypal structures organizing thought has remained in culture for a long period prior to even having an evolutionary scaffold to sustain it.

Underlying Myths

A specific evolved mythology underlies the Christian ethos and its primary text, the Bible. The Bible is written as a multi-author, frequently edited, strongly hyperlinked text. In Jordan Peterson's lecture series on the psychological significance of the biblical stories, he suggests that we ought to see the Biblical stories as morally significant in part because the Bible seems to have been strongly selected via the properties that guide Darwinian evolution (e.g. Dennett 2006):

1) Variation: The Bible varied over a long span of time and some parts were adopted while others discarded.

2) Heredity: Later versions of the stories of the bible recapture prior versions, and frequently the first appearance of a narrative in the bible turns out to be a modification of an archetypal myth that preceded it among the Greeks or the Egyptians (Peterson 2017-2018 Bible Lecture series).

3) Scarcity: The number of societies is finite, so is the number of individuals.

> Religious affiliation, in small scale as well as large scale societies is a scarce resource, thus there is competition for the seat of "fundamental axiological world explaining structure" in particular because religions often demand costly signaling of beliefs that are incompatible with the beliefs of different ideologies and religions (Purzycki et al 2016).

The Torah and Bible have been through this annealing and evolutionary combined process.

Other mechanisms less directly Darwinian in nature can also impact the promulgation of religious doctrine.

As Islam and more recently The Book of Mormon show, the control of moral fables in education of children is a powerful tool to socially engineer human behavior in adult life, even if fewer iterative processes that are selection like took place in crafting the texts and templates in question.

Secular Memeplexes are less filtered

In his magnum opus *A Theory of Justice*, philosopher John Raws extensively details how to construct a frame for a just society if you understand nothing about mathematics, evolutionary biology or sexual selection. In it, he proposes the usage of a veil of ignorance to judge between societies, where you would not know which individual you are. This is an interesting and valuable idea, as long as it takes into consideration both how those societies would progress through time as well as if the individual wearing the veil to make the judgment was allowed to consider how many individuals in such society are in each position. If that were the case, a society with one slave and a million happy families whose life is imbued with meaning would be preferable to a society with a million free servants who have not much to look up to or meaning and one king which they all serve. That judgment accords with intuition and with reflection. Rawls however invites us to ignore the numbers completely, and judge these societies solely based on which classes exist in them.

Society 1: Slaves and free people

Society 2: Free people and Kings

Since both societies have free people, the question to Rawls is whether it is best to be a king or a slave, and thus he concludes that the second society is best. That is the sense in which *A Theory of Justice* is a treatise on lack of quantitative knowledge.

But it also shows a remarkable ignorance of evolution, both cultural and biological. This is excused in part due to the book being written before the popularization of Hamilton's kin selection discovery later popularized by Dawkins (1971), and before most discoveries in evolutionary psychology.

Theories of Ethics, even the most famous one such as Rawls' or Parfit's *On What Matters* will often be subject to this potential failure more of crystallizing into an early error and not fixing it subsequently. As I see it, substantial damage was caused by Rawls' ethical problems in his theory of Justice, and those influenced by it.

We can attempt to improve if not fix Rawls with one of his predecessors, Bertrand Russell, who stated: *Both in heart and in mind, though time be real, to realize the unimportance of time is the gate to wisdom* (Russell, B. 2018).

Most global consequentialists, utilitarians (Greene 2013), and effective altruists (MacAskill 2015; Singer 2015) would agree. If future people have the same moral worth as present people, we should judge societies not only as they are in the present, but in how we expect them to evolve over time. More accurately, we should imagine the society running for however long it will, and wearing our veil of ignorance not only assume we don't know if we are the slave or the happy families, but also not assume we know if we are now or later. So a society of happy party people who raise their kids by leaving them alone in thrash and hoping for the best may not be ideal, even if there are no kids born yet, if compared to a slightly less joyful and party heavy society that takes care of future generations, leaves them a written legacy, functional institutions, procedural knowledge, healthy attachment, love, etc…

Sadly, the desperating biological science of human differences which progressed stupendously since Rawls' treaty came out has revealed dire news (Pinker 2004, Harris 2011, Bouchard et al 1990; De Menie & Jacobs 2013). First, most interventions don't really change human biology for the good much. To make people more moral, more emotionally capable, or smarter, is very hard with cultural intervention. It is very easy with nutrition and food, up to a point. But above that point, it is very hard again. Most of the interesting and relevant characteristics that differ between people (big 5 personality traits, intelligence, 10 aspect subdivision of the big 5, disgust and pain sensitivity, and happiness) seem to follow a pattern of being at least 50% heritable (and thus presumably biological) and 50% we don't really know, but we know for sure it is not school or parents (Pinker 2004, Harris 2011, Bouchard et al 1990). So to judge a society, it does not suffice to judge which roles different individuals occupy in it at a given time, but also which evolutionary patterns are likely to ensue in that society, and how will they change the biology, and therefore, the personality, of the generations to come.

It was good to find out that a lot of improvement could be had by increasing health and nutrition, but, for similar reasons, it was bad to find out that there's a hard limit on how much purchase those interventions accrue, and that cultural interventions have little to no effect (Dutton 2018).

If that wasn't bad news enough, the largest intelligence study to date, the China intelligence study (Yong 2013) as well as many others found that most of the characteristics we value and would like to improve in future generations are pleiotropic in nature (they are the byproduct of many genes interacting, not one or two genes that could be edited with CRISPR for instance). So the promise of genetic engineering as a morally tenable substitute for eugenics came out with a big empirical impediment making it much harder. Eugenics itself is pragmatically hard to execute, morally complicated almost regardless of which characteristics you choose to select for, and very time intensive as well, as it takes thereabouts of 25 years to form a human from scratch via natural means.

This leads us to the conclusion that it would be desirable for those interested in altruistic action to find cheap, effective, morally desirable, and socially accepted ways of improving the gene pool of future generations in a reliable manner. A tall order, given the history of the 20th century.

Lessons from the 20th century

Through the twentieth century some concerted efforts for finding a reliable way to create desirable societies backfired to the tune of over 100 million dead. To make matters worse from the perspective of an agent trying to increase altruism, multiple different styles of improvement of a society were tried, and multiple have failed. Germany attempted some versions of eugenic selection as well as segregation, isolation and expelling of particular groups. Russia and what later became the soviet block attempted an upheaval of the social order and hierarchies, and China as well as other Asian countries subsequently tried a different format of reorganizing the social order, markets, mating, and genomic assortment. In all these cases cited, the consequences were beyond devastating, involving genocide, forced labour, hyperinflation, social unrest, civil war and collapse of much of the social fabric of different regions or different populations within a country at a given time. Understandably we are now far more skeptical of revolutionary thought in politics as well as eugenic thought in the biosciences. This predilection is not evenly distributed across the world though. Some high G - the biological component that IQ attempts to measure - countries (Woodley & Figueiredo 2013) with high science production, notably China and Japan, are less cautious when discussing policy that involves biological change (Yong 2013). The great China intelligence study for instance attempted to detect genes more responsible for intelligence by collecting the genomes of many brilliant thinkers (Yong 2013) with the implicit purpose of improving the genetic component of intelligence in the nation. Also in China the one child policy was arguably the largest biological experiment in history, changing the incentive landscape such that families grew substantially smaller within a generation, and sex selective abortion led to gender gap of around 30-40million women, which of course has dramatic consequences for the cultural rituals and behaviors involving courtship, mating, and family formation.

The landscape of the biosciences has been substantially influenced by the conception of eugenics as well as debates about early stage development in most countries in Europe and the Americas. From an altruistic standpoint, it is necessary to balance bioethics and artificial ethics regulatory bodies and regulations in such a way that it isn't so strict we completely lose oversight and all the scientists flee, but also isn't so unrestricted that we risk making mistakes similar to those made politically or

scientifically in the 20th century, which led to some of the worst ethical catastrophes in recorded history. Some of those catastrophes are said to have been anticipated by one of history's great moral thinkers, though not necessarily one revered by altruists, Fredrick Nietzsche.

The Death of God and the Rise (Return) of Collectivism

Peterson interprets Nietzsche's claim that God is dead and his death shall lead to the death of millions — a claim of unusual foresight given the Holodomor and the Holocaust — as meaning that the seat of the explanatory framework oriented by Christianity, and thus, by the Christian God, would be taken over by other ideologies, perhaps less complete in their capability of guiding individuals through archetypal structures of living a good life. Peterson contends that might have led to the strong reemergence of collectivist ideologies such as Marxism, Nihilism (and their child postmodernism) and Nazism which fail to capture an aspect of the Christian narrative which had long evolved and that he considers paramount, the fundamental sacrality of the individual at the individual level, as imbued with a spark of divinity which, in its evolutionary predecessor Egypt, was restricted to the royal family, but in Christianity is present even among thieves, whores and social outcasts.

If we circle back to Ellen Clarke's discussion of the inter-level competition among levels of selection in the evolutionary struggle, what we begin to see here is Christianity as a prima facie apparently paradoxical form of collective individualism, where the unit of selection individual unites with the unit of selection Christians, to the detriment of all other levels of evolution and biocultural organization.

This focus on the individual as a place of sacred locus, as being valuable in himself, is, Peterson would argue, what makes Christian societies and those heavily influenced by Christianity less amenable to becoming governed by collectivists who are willing to sacrifice many individuals for the collective unit of selection, as observed in Bolshevism, Nazism and Maoism. The value of the individual being inscribed strongly in the Christian mythology would be the cultural force stopping the collective level from more completely dominating, as is the case with different units of selection such as Nazi germany, non-Sufi Islamism, and some varieties of Judaism.

When presenting this argument to other academics and independent researchers, I have often been faced with objections pertaining to subsets of monoethnic cultures that share a religion. That is, someone advanced the hypothesis that the individual alliance with the collective was not necessary to prevent collectivist sacrifice or genocide, using examples of non-religious monoethnic populations like the Chinese Han, or the Japanese.

One of the features of Christianity, Islam and Judaism have is that they seem to be religious structures that coevolved with constant influx (specially Christianity and Islam)

and outflux (specially Christianity and Judaism) of individuals. The structure of incentives surrounding a religion can have substantial effects in its mythological stories, behavioral propensities, and rules. So the objection that populations that have been isolated for millennia in islands or by mountains, where the average interaction is with someone who shares substantially larger than chance number of genes, and that via iterated crop raising selected out all but a few of the so-called warrior genes (Sapolsky 2017) is not relevant in a discussion of the evolutionary battles between major religious groups. In fact, eastern religions do not seem to be well adapted to being a minority religion, as Judaism is, nor for the constant influx and outflux. When populations spread in migratory groups from China and Japan to the most diverse parts of the world, they seldom sustained any connection with the major religions of China and Japan. Even with the rise in the West of so-called Uptown Buddhism and related mysticisms, there doesn't seem to be a sufficiently distinct behavioral costly signaling border such that endogamous mating with co-religionaires would become common. So even though China and Japan could be considered superorganisms as a coalition of the State and people, I would not judge these coalitions to be primarily religious, but instead ethnic and geographical. Further, Japan partook in some of the worst acts against individuals during the second world war (e.g. unit 731), and China conducted experiments that led to the death of more people than any other individual human conducted catastrophe in history. Thus, if anything, they reinforce the hypothesis that absent a religious structure incentivizing a two level alliance between the religious group and the individual, individuals are more likely to become disregarded in favor of the group level.

The level of adaptive flexibility necessary to adapt to multiethnic, high inequality, large, permeable societies has not been achieved by religious structures that didn't geographically, in one or another way, move between different biogeographies, and dealt with influx and outflux of individuals from different societies. We can call more geographically determined superorganisms, such as China, Korea and Japan, as allotropic, for our purposes. For the reasons given above, allotropic superorganisms do not constitute an epistemic menace to my analysis.

How do we connect these macroscopic considerations with our original question of scalability of altruism in space and time? In this case it is easier to see the connection because nations, as well as religions, are scaling structures (a nation doesn't change what nation it is if its population doubles, for example). So any coalition, superorganism, or macrostructure which elicits or promotes altruism at a superorganism level can be favorable to scaling, though it doesn't necessarily have to. Superorganisms are, by definition, larger and more scaled than individuals or families, and they can scale to very large numbers, as is the case with Islam, China, and Christianity.

Flexibility, Cranes and Generativity

Not only at the supra-level of "Christianity" "Islam" "Judaism" and "China" for example do these superorganisms partake in evolutionary battles.

In 1978 Gerald Edelman suggested the since recurring idea of neural Darwinism, suggesting that coalitions of neurons fight for domination in our conscious, and attended to, awareness, and there is a Darwinian process of sorts. Likewise, synaptic culling has been suggested to organize itself by a combination of Hebbian and Darwinian processes. Even Feral Neurons have been defended by Dennett 2017 to be fending for themselves in Darwinian ways. The metaphors of Darwinian battles in substructures of our organisms range far and wide.

Likewise, I contend with human religions, ethnogroups, ethnoreligions etc… There is often competition between different structures intra and inter-levels.

People don't have ideas, ideas have people. - Jung

A scholar is just a library's way of making another library. - Dennett

So Mormons and Catholics compete for evolutionary primacy, not necessarily consciously, but through different strategies of converting and raising humans, and accumulating and distributing power and wealth, which produce different genetic and cultural incentives as well as different levels of responsiveness to environmental variance and change.

In *Darwin's Dangerous Idea*, Dennett puts forth a differentiation between Cranes and Skyhooks in the evolutionary process. Cranes build on lower, simpler layers, accelerating the speed of some aspect of the evolutionary process (e.g. mutation generation and testing speed). Skyhooks, like Munchausen's bootstrapping, or intelligent design, are miraculous forces that somehow would pull the evolutionary process upwards, a concept mostly developed to counter simple Intelligent Design theories.

The USA's constitution and initial social organization can be said to be a Crane for the multilevel competition between superorganisms. The religious freedom and vast expanse of land enabled the acceleration of the generate and test process of many religious branches in the evolutionary tree, from Hutterites and Mormons to Evangelicals and Catholics. It also worked as a Crane due to being a safe haven to some religious groups persecuted throughout history, most notably European Jews in the pre World War II era.

Is there a lever of true biological altruism we can work with?

Multilevel Selection

"Selfishness beats altruism within groups. Altruistic groups beat selfish groups. Everything else is commentary" - Wilson & Wilson 2006

Historically, the paradox of altruism in biology was "how is it possible that selfish individuals evolve altruistic behaviors and stabilize them in a population?" Despite the selfish nature of genes (Dawkins 1999) and other units of Darwinian transmission (Jablonka & Lamb, 2007), altruism at the individual level (cost to self for benefit to others) *can* and *does* arise because of several intertwined factors.

To a first approximation, group selection, as a subtype of multi-level selection; superorganism selection; somatic cells selection; species selection, and individual selection - only happens when *the selective forces operating on that level coincide* with the allele's fitness increasing in relation to all the competing alleles.

Alleles, epigenetics, and learning can program individuals to be cooperative if agents "expect" (consciously or not) the interaction with another individual, say, Malou, to: (a) Begin a cycle of reciprocation with Malou in the future whose benefit exceeds the current cost being paid; (b) Counterfactually increase their reputation with sufficiently many individuals that those will award more benefit than current cost; (c) Avoid being punished by third parties; (d) Conform to, or help enforce, by setting an example, social norms and rules upon which selection pressures act (Tomasello 2005). A key notion in all these mechanisms based on this encoded "expectation" is that uncertainty must be present. In the absence of uncertainty, an agent in a prisoner's-dilemma-like interaction would be required to defect instead of cooperating from round one, predicting the backwards-in-time cascade of defection from whichever was the last round of interaction, in which by definition cooperating is worse. The problems that in decision theory people tried to solve by conceiving of the complex *Timeless Decision Theory* (Yudkowsky 2011) and theories developed based on it, evolution solved by *inserting stupidity*! More precisely by embracing higher level uncertainty about how many future interactions will there be.

Finally, altruism only poses paradoxes of the group selection kind when we are trying to explain *why a replicator that codes for Altruism emerged? And* we are trying to explain it *at that replicator level*. It is no mystery why a composition of the phenotypic effects of a gene (replicator) and two memes (attractor-replicators) (Henrich et al 2008) in all individuals who possess the three of them makes them altruistic, if it does. The selfish nature of the constituents does not imply that the combination of them would be selfish - that would be a mereological fallacy. If we trust Jablonka & Lamb (2007), there are four streams of heredity flowing concomitantly: Genetic, Epigenetic, Niche

Construction, and Cultural. Some of the hereditary entities flowing through evolution's cascades are not even attractor-replicators (niche construction for instance), they don't exhibit replicator dynamics and any altruism that spreads through them requires no special explanation at all!

Altruism and Evolutionary transitions

Evolutionary forces sculpted most of the design of the world we see around us. Whether biological, cultural, or economic, forces have arranged and rearranged the world around us producing entities that populate the biosphere. This process is mediated by evolutionary transitions (Smith & Szathmary 1997, Clarke 2014) where a set that was previously composed of separate beings becomes a unified entity, for example many cells become a pluricellular organism. Evolutionary transitions are relevant to the study of altruism for two reasons:

1) They are necessarily preceded by synergistic interaction;
2) They are frequently preceded by cooperative or altruistic interaction between the constituent entities which later become altruistic.

Relaxation, drift, transgenerational altruism in the form of niche construction, and error can all contribute to the process by which a set of altruistically interacting or cooperating agents progressively loses autonomy. That, as Hui and Deacon (2010) suggest, may initially lead to a process of addiction to those interactions, a self-domestication of sorts, where autonomy is lost at the individual level and substituted by these interactions. At the limit, however, loss of autonomy might lead to full individuality being constituted by what originally were separate beings. The most famous example of this process taking place is the emergence of pluricelular life (Szamáthary 1997). Some amoeba groups, such as the slime mold, can dynamically enter this interstate between unicellular and pluricellular.

Evolutionary transitions can turn many into one, and altruism requires more than one. So evolutionary transitions can spell the end of an altruism cycle. They may however also spell the beginning of an altruism cycle one level higher (Clarke 2013), as the ex-groups, now individuals, can also form the same type of cooperative or altruistic relationships that were previously possible at a lower level.

Deacon and Clarke on transitions from group to individual (and sometimes back)

Terrence Deacon and Ellen Clarke have both worked in questions about so-called evolutionary transitions that are of our interest here. What is distinctive about their

work, in particular hers, is that it investigates a larger plethora of questions than the traditional approach of asking solely about "what are the conditions that enable an evolutionary transition, once it happens, to become stable, and, in particular, how do larger groups prevent the emergence of evolutionary competition at lower levels?" This is a fascinating question, and due attention was given to it in the literature, but many more lines of inquiry regarding transitions can bear fruit. Let us examine them in turn.

How do new levels emerge?

The question Deacon is interested in is related to the emergence of new levels of complexity. An evolutionary transition is a case of such emergence of a cluster of constraints that cannot be exhaustively understood looking only at the lower levels of resolution. Deacon is interested in the *how* question. He studies the emergence of discontinuous levels in many different domains: from symbolic cognition which he argues is discontinuous between us and other animals (Deacon 1998); to phylogenetic semiosis, which he argues started alongside life and "ententionality" (Deacon 2011), a wider form of intentionality; to semiosis proper, where he decomposed the Pearcean iconic indexical system into a cognitive process. The question of emergence of new levels of social organization bears many resemblances to these other hierarchical emergences, and Deacon characterizes the social-specific mode of transition thusly:

At first we have a general evolutionary principle in which there is duplication of an entity, this can be an individual, a cluster of genes or even a function performed by some member of that society. Once the duplication occurs, there comes to be some degree of redundancy. That is, the function, now performed twofold, may be needed only once, or some in-between quantity. That leads the way to a process of degradation, where part of the functionality of the "crisp" original may be lost, as its twin, or slightly degenerate twin can now account for the loss of function arising from degradation. This, in particular in sexual or cultural evolution (as opposed to asexual reproduction) leads to a new state of recombination. The newly degraded now mutually interdependent entities may start to acquire not only different properties in virtue of the degradation, but also synergistic properties. Properties that allow them to divide labor in a way more effective than two copies of the original would, for example. This, he contends, should happen in social semiotic processes as well as in biology.

Higher order human social and semiotic units evolve in the context of loss of autonomy of lower level units due to hierarchic transitions of social organization, and I can anticipate a similar process to take place at the level of semiotic process and signification. What would this look like? At the level of word meanings and symbolic reference, we observe a process of cognitive duplication of the structures that embed

reference, followed by a degeneration of the indexical component of reference in the second "layer" of cognitive representation of entities. This leaves them in a "free floating" state where the relations of indexicality and adjacency are now mostly internal relations to each other. That, in turn, allows symbolic manipulation and the symbolic process that is typical of our cognition, and that differentiates us from the other animals.

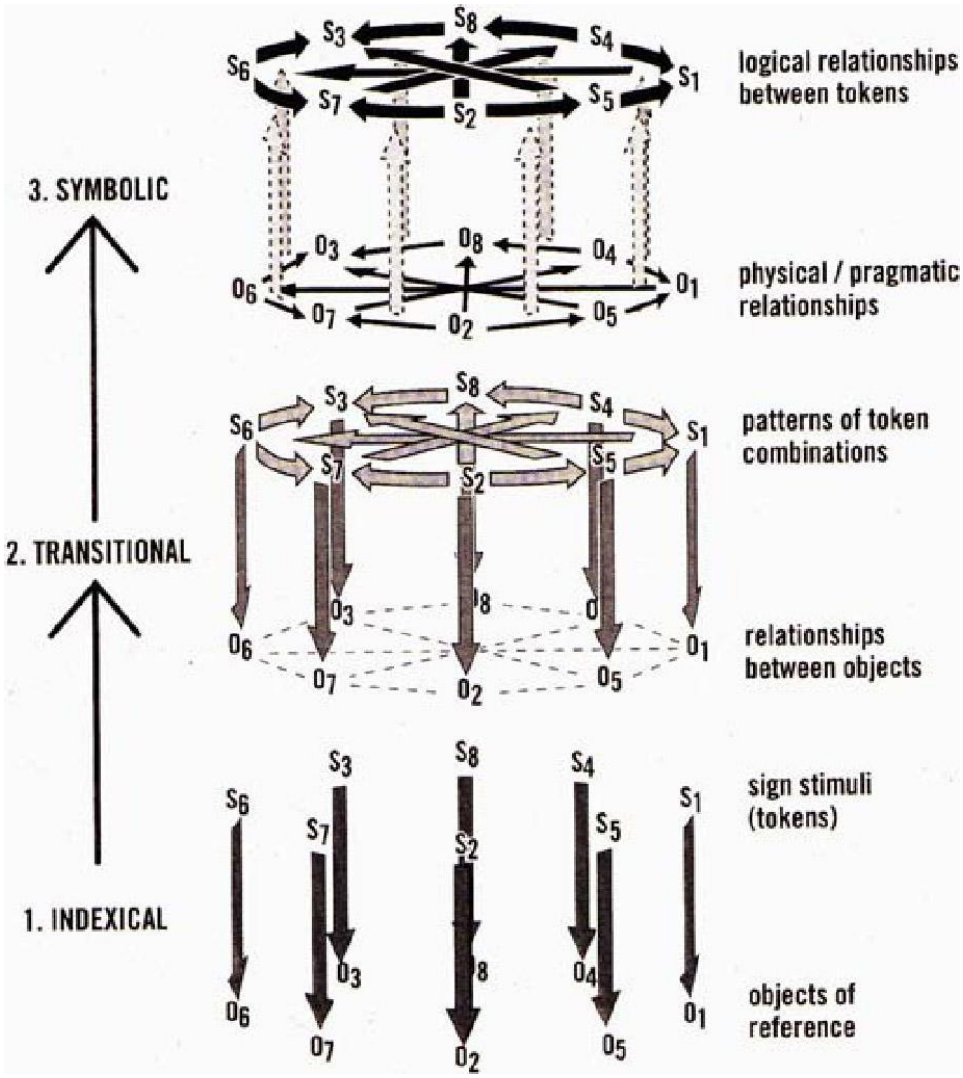

Fig 3; Symbolic cognition.
Imported with permission from The Symbolic Species, Terrence Deacon 1997

One of these symbols that starts off as an index and progressively evolves into a fully formed post degenerate symbol which can be manipulated is our notion of ourselves, our model of self. To illustrate this degenerative process, consider that while the reference for our identity and sense of self is still indexical, say, in children younger than 4, some levels of delusion in one's concept of self will be effectively impossible. The conceptual apparatus and semiotic cognition necessary to even entertain some ideas is

simply lacking (perhaps self-reference, first person subjunctives, and other theory of mind laden second order logic symbols would be a case in point). Later in adulthood, when not only our sense of self is well formed but also our sense of other people's selves is already formed, there is the possibility for manipulating which of those symbols goes where, allowing some people to enter delusional states such as believing they are Napoleon or Jesus, that is trading the position of symbols for individuals for one's own self representation.

Human representation of selfhood is partially internal and partially external, as in an extended mind (Chalmers & Clark 1998). To some extent this is due to semiotic process of offloading the umwelt into regularities of one's environment, thus leading an organism to experience less of the world and the world to be a pattern either in the individuals predictive neurocognitive mechanisms (Hawkins 2007) or in the regularities of the environment themselves securing functional behavior without semiotic representation. Benign ignorance.

The individuality itself of our ancestors likely was partially offloaded to and constituted by commitment contracts and the expectation of others that one would continue to be the same over time (Adorno 1947, Trivers 2011), that is, it was offloaded in part into the social realm. Just like parts of our minds are transferred into the ecological, social and technological Umwelts, our individuality was partially offloaded into social connectives. The transition of state one to state two organisms - In Clarke's definition discussed earlier) is not without friction. At every level inside complicated groups of organisms that experience some group level selection, there are still entities whose incentives are to act against the interests of the larger whole of which they are part. Individuals have incentives to act against groups and groups might develop counter-strategies to prevent individual defection, for instance the scapegoating mechanism (Girard 2007), or a "moral" system based of third party dynamic coordination to reduce friction to the group at the cost both to individuals and to morality (DeScioli 2013). Within cells there are also battles between the incentives of different parts, possibly the most famous of which being the one that lead to the emergence of sexual reproduction, where female organelles defeated male organelles, becoming the sole providers of mitochondria and other subcellular components for the next generation (Ridley 1994).

It is in this context that we can think of evolutionary battles not between organisms, but between levels, and in this context that the tension between the individual and other intra-individual and extra-individual levels of selection becomes more clear. This orthogonal dimension of evolutionary friction is where the meat of the tension between individuals and groups, and individuals and their subcomponents (Dennett 2017, Simler 2013) lies. One way for altruism to scale is for larger and larger levels of selection to emerge over time, by growth of previous levels, mergers through

synergistic interaction, or de novo emergence.

Multilevel selection in four dimensions

Freewheeling evolution is ubiquitous. The conceptual space of evolutionary change has been suggestively divided (Jablonka 2007) in four dimensions (4D): genetic, epigenetic, niche construction and cultural evolution.

Orthogonal to this partition, we can partition the space of evolving entities in different resolutions of selection, we can call those levels (e.g. with *gene, individual, group selection* as levels in the genetic dimension) corresponding roughly to how many constituent entities of a lower level are best grouped together to provide an effective causal account of evolutionary development. All undergo drift, transmission mistakes (noise) as well as selection pressures enabling and enforcing change at that level.

Here I propose the *4d evolutionary matrix* model in Figure 2 (below)

	Minimal unit of selection	Low level	Mid-level	High-level	Very-high-level	Global Level.
Cultural-Evolution	Meme	Song, book, blog post, idea.	ideology	state, corporation	scientific method, economic systems	Singleton
Niche-Construction	Location	Ant-trails, simple tools	Buildings, parks.	Bridges, borders,	Agriculture.	Global Warming
Epigenetic	Multiple Nucleosomes (Robinson 2011)	Chromatin, RNA, Proteins, everything small inside an egg cell	Protein chain activation clusters	Mother-child methyl-group diet based protein chain activation.	Doesn't exist	Doesn't exist.
Genetic	Gene	Individual	Group	Coalition	Super-group	Everett branch (with all inhabitants)
4D Mix	Doesn't exist	Gene-epigen combo that codes for a beaver dam design type?	DRD4 r7 allele plus epigenetic failures that encode high probability of hierarchical individualistic thinking.	Religious population /sect, monoethnic nation	Religion, civic nation	

Figure 4: A tentative representation of the evolutionary complexity matrix, with the 4 dimensions in Jablonka (2007) divided into six analogous levels. Notice not all levels of complexity are reached by all dimensions.

Though some entities evolving in these do interfere with the pace and cadence of the other levels, they do not stop them - e.g. medicine doesn't curtail evolutionary

modification in the immune system, it simply alters the fitness landscape, relaxing the relative evolutionary pressure traction on that trait.

It is noteworthy that in sexual animals there's a tradeoff between evolutionary pressure that pertains to natural selection, versus sexual selection. This is because those two constraining forces often operate in conflict - for instance in costly signaling. Abundance of resources may lead to iterated sexual selection for characteristics that reduce individual survivability, but still increase the number of long term offspring, as in the case of the peacock's tail or gangsters (Daly 2017).

Here I accompany Jaegwon Kim and Judea Pearl (Kim 1982, Pearl 2009) in saying that selection is happening at a particular level when, at that level, it most precisely determines and predicts events happening at any level in the future. This notion is related to the Price equation and to the discussion in Clarke (2014) which suggests a methodology that permits determining levels of complexity where selection is stronger or weaker.

Evolutionary forces act in at least four dimensions (Jablonka et al 2007), and within these, at many different levels of complexity (Wilson & Wilson 2006).

Suppose we want to represent an evolutionary cladogram in another dimension besides the branchings for which cladograms are known. In particular we want to represent the transitions where the level at which natural selection is exerting most force (this can be determined e.g. via price equation and modifications thereof). We could then color code different levels of constraint, and use intermediate colors for cases where selection is still in transition. For instance, the appearance of superorganisms would be represented by the emergence of a blue cloud from a green one, and slime mold collectives would be an in between color. In this case the colors determine the organism level, such that for instance organisms that are in between state one objects and state two objects such as slime mold would be able to morph color, if the representation was instantaneous and dynamic, or be an intermediate color, if it was lower temporal resolution.

Relations that at one level are seen as altruistic, become organic at a higher level (Clarke 2014). Different organs of the same organism, so an extra dimension (color for instance) can be used to represent, within evolutionary cladograms, the existence of fitness constraints that are operating at a higher level of selection.

The level at which evolutionary pressure is being exerted is one possible proxy for individuality, it is evolutionary individuality, or fitness individuality. And different modes of individuality conflict with levels of selection to different degrees, but I argue that all of them are in conflict with one or another infra-individual or supra-individual force, be it biological (Sober & Wilson 1998; Buss L. W 2014), cognitive (Blackmore 2002), moral (Greene 2014), personal identity (Ainslie 2001), economic (Sowell 2014) or sometimes even energetic (Deacon 2011). The rise of altruism and cooperation gives primacy, by

definition, to the supra individual level, and this primacy cannot always be made consistent with individual primacy.

Exceptions and Unusual Cases

Although the regularity of a negative correlation is undeniable, it is not necessary, there are cases where altruistic interactions actually work in favor of stronger individuality.

Take the family unit: although the unit is a group of individuals that are unusually altruistic towards one another, the existence of separate roles for the different individuals in a family actually strengthens the group, and their difference makes them more differentiated, more distinct, and thus, more individualized. The loss of autonomy is accompanied by a gain in narrative gravity, and these additional distinctions make agents more individualized. The formation of narrative archetypes which represent different parts of the family structure (Peterson 1999) is constitutive of personality, which is one of the animal specific types of individuality, particularly salient in humans.

Another case also dependent on sophisticated cognition is the reputational tracking cognitive machinery. If your social group altruistically tracks the reputation of individual members and prevents defection, this strengthens the incentives for individuals to be cohesive over time, in order to secure accountability and, in the human case, actual and perceived moral integrity, the appearance of being a monolithic agent taking predictable reliable actions over time.

Robert Putnam argued (2000) that the loss of prosocial civic engagement has deteriorated individuality among Americans, suggesting that individuality took a hit from the decrease of in-person social intercourse in American life.

Guilt based morality cultures - as opposed to shame based - have high prevalence of altruism (e.g. American protestants donate to charity) while sustaining individualistic values (e.g. emulating Christ as a schelling ideal individual, personal responsibility, individual salvation etc...), which seems to be an anomaly particular of Christian cultures and groups descendant bioculturally from 16th century catholics who prevented cousin mating within church members, fostering a reduction in clannishness and an individualistic rise in cultural norms (Henrich 2016).

Lastly, only among humans, the moral circuitry designed for moral punishment (DeScioli 2013) seems to be designed not to maximize moral integrity, but to dynamically reduce the third party cost of conflict for any conflict.

Altruism: Fast and Slow

System 1 and System 2 altruism

In *Thinking, Fast and Slow* (2011) Daniel Kahneman brings us into a trip into the world of experimental individual psychology, and one of its most robust predictions about how we think. It argues we have two main cognitive modes, one, composed of quick heuristics and fast responses which are in some sense pre-loaded in our behavioral repertoire. We are often required to react fast so it would make sense that evolution would end up endowing us with a fast system of reaction with quick efficient responses to salient features of the environment.

In fact charity scientists examined the magnitude of donation to increase if someone is asked to bear cognitive load (keeping a number in mind). Altruism can be caused at times by inability to think!

Dual Process Theory: I take my models of Dual Process Theory from Greene (2013) and Kahneman (2012), and focus in particular on the interaction between Greene's description of point-and-shoot, fast, automatic morality and the work done by John Mikhail (2007a; 2007b; 2007c; 2011) on the idea of a universal moral grammar as well as the mismatch between the moral cognitive process and the judgment of moral actions.

In our process of judging a moral action such as a trolley problem (Mikhail 2007) can be interpreted to imply we utilize an Alpha-Beta pruning algorithm to determine the moral value of actions up to the point where these action events are framed as *means*, and our pruning algorithm becomes defective (in predictable ways) in branches where events are framed as *side effects* but not when they are framed as *goals*. That is we continue to memorize the value of consequences that belong to a causal chain where action A is means to B is means to C is means to D, but we fail to compute in our moral calculus side effects, such as when A is means to B and has side effect Z, B is means to C. We compute the moral value of A, B and C, but forget to compute Z.

This hypothesis is consistent with the data about temporal sequence and causal structure from moral cognitive neuroscience.

System 1: Grateful, kind and caring.

System 1 is more empathetic, kind and nice. It considers the concrete other as presented immediately, and reacts to it swiftly and what we perceive as unthinkingly (another name for unconscious cognition). So a priory, this automatic fast mode would seem to be the more altruistic of our two psychological sides.

Who is more altruistic, system1 or system 2?

Yet this is not the case. The distribution of altruistic actions is very likely fat tailed (Kokotajlo 2020) and thus taking the time to find a few highly impactful actions would, in an aggregative consequentialist framework, completely dominate a series of local empathy based actions, even if they were far more numerous. A similar argument is made in *Against Empathy* (Bloom 2017). Our natural tendency to cooperate is unable to compute information that never benefited our kin nor was taught to us in school or anywhere. In a globalized world it is much easier to create substantial benefit to others by using the manual, slow mode of the mind, and through deliberation to decide to allocate resources.

How can we utilize these two systems to manipulate ourselves into being most altruistic?

Both of the systems can be hacked to increase one's level of altruism though. Many groups who desire us to deliberate towards them as we would our close kin call oneselves brothers, sisters, and other familial terms. This is an attempt to use our tendency to favor kin to favor those who share fewer genes. It is technically a manipulation (Dawkins 1971). Similarly, photography and video can be used to give a strong immediate felt sense which may incentivize a particular type of altruistic action. VR cameras in war zones have been used to try to get people's system 1 to empathize with the victims of war, for example.

System 2: Just, benevolent and calculating

Less kind and caring than its numerical predecessor, System 2 has reached a higher level of moral development in Kohlberg's scale. The notions of justice and the process of calculation participate in our deliberate, long term, decision making process in ways they do not in the quick trigger one. System 2 is more impartial, and able to distribute resources in a way that, ex post, we would usually deem more morally desirable on reflection.

Psychopathic autistic altruists

No contradiction despite appearances. A consequence of these modality disparities in our altruistic behavior is that some aspects of some mental illness or personality disorder can, in some cases, facilitate altruism. A military general, statesman, or CEO might at times have to make decisions which would be considered

non-empathetic, cold and calculating from the outside, that would be impossible to take if they felt normal levels of empathy, yet that still would benefit the largest number the most. In part, that is why psychopaths are 3x overrepresented among CEOS. Part of it is that they would also be able to make the non-empathetic selfish, non altruistic decision. Autists, likewise, overrepresented both in philosophy departments and among effective altruists, are also able to perform actions with consequences that socially normal people would not. At least in some cases, this provides a relative advantage in altruistic capability.

Me versus us

Joshua Greene (2013) distinguishes two useful modes through which we conceptualize morality, and associates Me versus Us with System 1, automatic morality. He contends that we have evolved long enough in situations where there are moral conflicts of this type that our brains are already decently attuned to it. We already have the specialized circuitry to undergo that type of reasoning as it were. We can trust our instinct when it comes to dilemmas of Me versus Us.

Us versus Them: the tragedy of common sense morality

Yet, he suggests we have not evolved the cognitive capability to think of Us versus Them - much of the anthropological record of Sonté would favor his hypothesis - We don't have the natural, fast sped, inclination to solve moral problems with conflicting values between those values we share with those in our society and a different one, our moral tribe and another moral tribe. He calls this problem the tragedy of common sense morality. In short, what we mean by common sense stops making sense at the boundary of our moral tribe. His book is an attempt at offering an alternative for how to think through these types of problems.

What got us here won't get us there 1: kicking the ladder

What are altruists to do in this case? The suggestion Joshua makes in *Moral Tribes* is to use utilitarianism as a currency for cases where the allocation of resources is undecided. Figure out what would make the most most happy, and allocate the newly found common pastures accordingly. It's an elegant solution.

Whether we implement Joshua's solution or not, the meta point is that what got us to the level of moral development we have achieved thus far is a set of cognitive and psychological evolved mechanisms that cannot bring us to - literally - the next level. As Sloan Wilson and E O Wilson (2006) would contend, we evolved our cooperation through aggression and competition between groups. So we lack the cognitive structure

that would cause us to go one level up automatically and be as cooperative with a group of others as we are, for instance with our teammates in sport or war. We have, in other words, selfish reflexes.

Selfish Reflexes: the threat of Christmas past

There are strong attractors that pull our cognition into less altruistic modes of thinking. System 1, empathy itself, genetic drives, tribalism, automatic racism, and many other adaptations. If we are to circumvent these threats of Christmas past in order to increase the amount of system 2, manual, slow, altruism and cooperation we are able to engage in, this will require an ergodic process. That is an energy consuming process of teaching or otherwise spreading values, algorithms and ideas that entice, facilitate, or otherwise cause altruistic behavior. Reputational systems in Ebay and Uber are examples of systems that facilitate cooperative behavior. If we are to seek altruism that is as global as possible, we should also distinguish two kinds of systems that facilitate cooperation. Sortition based processes, and enhancements. A sortition process makes it easier to detect who is or is not likely to cooperate or act altruistically, but doesn't make a given individual more or less altruistic. An enhancement actually improves their level of altruism (Fabiano, J. 2021).

Local altruism

The strongest drive that can supersede an aggregative approach to altruism and cooperation such as the one I have been tacitly advocating so far is the drive for local altruism. Instead of considering what is the most benefit per resource spent anywhere anytime, we normally only consider the immediate factors in a situation: should I give this beggar asking for money now, some money now, here? To expand the scope of possible actions increases the probability of accessing the fat tails of altruism where the impact can be orders of magnitude larger (Kokotajlo 2020). So abandoning local cognition is in itself an altruistic intervention, in expectation.

Religion and altruism: there is good without God

Some people associate the notion of altruism with churches and their specific type of local charity (Wilson 2012). Just like churches are evolved structures which created an adaptation through which they are able to allocate some resources to altruistic actions, likewise our brains and us are also evolved structures navigating the intersection of culture and biology and thus are able to mobilize resources and actions in altruistic ways

without having to rely on any particular organized religious group or system to do so. Religion can help tap into the multi-level selection cognition, and cause us to think that an altruistic action is for the good of the group, the same way calling someone 'sister' or 'brother' can. A punishing God that punishes non-altruists (Henrich 2013; Sharriff & Norenzayan 2011) can also be a useful conceptual abstraction incentive to cause altruistic behavior.

Abstract Religious Altruism and a Collective Identity

Two aspects are worthy of note when it comes to religious altruism, the fact that it is often parochial - directed at other members of the same religious group - and the fact that it substitutes the almost as simulacrum, the genetic proclivity to prosociality with a religious version. Religious groups frequently denote this symbolic substitution with fraternal and paternal titles and roles given to church members, mimicking the social relations of small groups or families. Social regulation mechanisms evolved differently for groups of different scales. The time necessary to rally someone on behalf of a God or a nation is higher than that required to rally a family member to defend their kin. Having larger Gods, closer to omnipotence and univocity, is an effective strategy to coordinate increasing numbers of individuals, as demonstrated in the victories of monotheisms the world around, in terms of the number of people persuaded by those ideas. Religion also helps regulate order and mating of those who subscribe to the collective identity (Sharriff & Norenzayan 2011). In a highly religious community the benefits of partaking in religious ritual and professing religious belief could be high in terms of status, probability of passing your genes on, mate quality - which itself is higher in religious people who bear lower mutation rate - and the costs of not participating can go as far as ostracism or death.

Lastly a collective identity and religion can be seen as a set of rules and constraints which is being followed by the religious participants, and in doing so, a niche environment that allows for optimization, similar to how the legal system operates as an abstract scaffold that enables individuals, corporations and groups to anticipate and predict how the future will react to their actions.

International Relations

"There can be no doubt that a tribe including many members who, from possessing in a high degree the spirit of patriotism, fidelity, obedience, courage, and sympathy, were always ready to give aid to each other and to sacrifice themselves for the common good, would be victorious over most other tribes; and this would be natural selection." — *Descent of Man* p. 166 - Darwin

"If you don't have some kind of global cooperation, nationalism just is not on the right level to tackle the problems." - Yuval Harari

Large abstract entities like nations and countries can also partake in abstract action sequences such as cooperating, defecting, attacking, helping, and being altruistic towards other nations or other lower level entities like cities and people.

Some, like Harari (2020), Bostrom (2003), Zizek (2018) and Sloan Wilson (2020) believe if we try to focus our attention and evolutionary pressures at the level of nation states, this will not cause desirable future outcomes. Harari and Zizek advocate a strategy of increasing global cooperation. Bostrom argues that only a Singleton, an authoritarian entity so powerful it can suppress evolution at all lower levels, could possibly stabilize the aspects of our lives that we value without an evolutionary race to the bottom. Sloan Wilson (2020) argues that we have to push our mental modules for group favoritism up as high as possible, using e.g. Oström's famous principles for successful collectives, but going beyond that and trying to directly act in ways that are globally beneficial.

In international relations, the Realist school suggests that over time the balance of power worldwide will be basically US plus most Asian countries (Russia inclusive) against China, as that is what the "selfish at the country level" incentives predict to be the equilibrium state. Since this prediction was made long ago the US has indeed increased alliances with Japan, South Korea, North Korea and Russia, giving some epistemological substance to the Realists.

The Securitization school contends that labels can be used to weaponize countries. Transforming something previously ignored into "a question of national security" can considerably shift how the public and other countries see an issue, and in performing these semantic shifts, the perceived and actual power disparities between countries can be changed.

Authoritarianism has decreased almost monotonically as a percent of humans governed under it for nearly a hundred years (Norberg 2017)before a recent tipping of balance in the other direction, contradicting Fukuyama's famous *The End of History and the Last Man* blunder and showing that the future of international relation and forms of government is all but certain. China moved to a lifetime presidency. Russia and other ex-soviet countries moved towards longer incumbencies and more centralized power.

"It is true that political leader's ability to good is limited, but their ability to do harm is unlimited!" - Yuval Harari 2017

The scope of individual human altruistic action in ways that change the balance of international relations is narrow. Still due to the brittle nature of the incentives that govern market and political incentives between nations, as well as the amplifying effects

of nuclear warheads and artificial intelligence based attacks, the relative power of individuals to change the international balance of power has increased. Donald Trump evidently changed part of the balance of power in Asia compared to the counterfactual. Julian Assange might not have substantially changed politics in the ways Trump did, but he showed the magnitude of impact that a single individual can have, and this is only increasing over time. Fewer and fewer people wield a larger destructive capability. The number of people invested needed for a catastrophic risk (50 million deaths) or an existential risk (permanently curtailing humanity's potential) has been subject of academic study by the Uehiro bioethic center in Oxford 2013 (personal conversation, I was there).

An altruist can find leverage in the international relation scenario in two ways: attempting to acquire some power themselves, like Trump and Assange did, or changing the technological dynamics in ways that favor some countries and large agents more than others, which is arguably being done, for example, by TikTok - though not necessarily altruistically. We will return to the question of nationalisms when talking about superoganisms and types of cultural groups.

The decentralization of everything and how it incentivizes altruism

Centralization, historically, ebbs and flows. There have been periods of concentration of power into higher level entities (formation of Germany, reunification of Germany, formation of Italy, European union) and decentralization into smaller, lower level entities (Brexit, Balkanization, move towards local manufacturing during the Covid pandemic, etc...).

Within technology, the 90s and 2000s saw a period of decentralization (even stronger if you consider TV the precursor of the internet), then we had the last decade with a strong motion towards centralization in a few giants (Amazon, Facebook, Google, Twitter) and one strong decentralizing motion (Cryptocurrencies, in particular Bitcoin).

In the years in which I have been writing this (2015-2021) there was a strong decentralizing set of forces. Over a billion people moved from passive watchers and readers to commenters, tweeters, and sometimes content creators. Politically, there has also been a move towards decentralization, with the rightwing parties winning elections in many countries where that seemed surprising in 2015, including Brazil and USA, where I live.

Charity has also become more decentralized. There are now multiple types of charities, disconnected from the original venues where they were most commonly found in the past (State and Church) (MacAskill 2015).

Cryptocurrencies made it impossible for governments to tax the movement of individual capital (Ammous 2018). The energy matrix of wealthy countries also became

decentralized, with solar panels multiplying significantly in quantity, increasing the robustness of the energy grid.

Reputational systems such as Ebay, Uber and Amazon sellers also created a de facto substitute for reputational trust. We now believe the safe thing to do is let our 16 year old daughter enter the car of a stranger foreigner to go to a party and drink, not to drive her own car. This would be unthinkable a few decades ago. These reputation systems are very effective, and substantially increase the altruism and cooperation in the World. An altruist of technical inclination would be hard pressed to find a more effective use of their time for altruistic maximizing than finding more domains besides driving, housing, and trading where a digital reputational substitute for memory based reputation can be used.

While tracking reputations and identity do indeed facilitate some kinds of altruistic and cooperative action, other actions are helped by the precise inverse: privacy and concealment. The Internet allowed anonymous communication which can bypass censorship mechanisms even from the most powerful agent in the world, the USA armed forces. This created the conditions for many movements and changes which might retroactively be considered altruistic, from parts of the Arab spring, to the creation of Ethereum based cat token economies. Anonymous webpages enabled donations, crowdfunding for charitable causes, and an encyclopedia much larger than could be created in the world of atoms. The decentralization of money supply through cryptocurrencies could turn out to be altruistic if it deters sufficiently many governments from predatory inflation, and it could be deleterious if it facilitates criminal actions by authoritarian dictatorial regimes that would otherwise not be possible. It remains to be determined.

Besides decentralized currency, smart contracts enable infinitely many other trust invariant operations, which could become a more substantial fraction of the economy in the coming decades. A bridge between these technologies and the current mechanisms of trust performing similar functions would enable a deepening of verification without trust. Systems where you can verify cooperation even with an avowed deadly enemy.

Prediction Markets for sport and politics are somewhat common but not very predictive due to betting limits. They would also count as a decentralized decision process which helps accurately predict the future. As the volume of operations increases, they become a new decentralized prediction tool, similar in some ways to stocks, but available to a larger cohort of betters. That would be a decentralization of information in time.

From an altruistic standpoint, centralization and decentralization, per se, are neutral. Both have benefits and costs in different contexts. The USA individualism is a paradigm example of decentralized *politi,* whereas the Chinese centralized government is an example of centralized governance. China removed twice as many people as live in

the USA from poverty. In part due to their high average IQ (technically G which is an improved measure of intelligence, and more biologically laden) (Woodley 2012), but in part also due to top down centralized planning and organizing. The USA did similar if not more incredible feats throughout the last 100 years as the World's superpower. The standpoint of an aggregative altruist here, it seems to me, should be to desire that there is some variation between nation states so that the benefits of scale can be accrued by the whole world, while the benefits of decentralization can also be obtained, if not by all, by as many as possible.

Again the highest impact area for an altruist agent in international relations might be to prevent the tipping of balance of power, or to cause it, through technological change. Stabilizing a unipolar equilibrium is a surefire way of increasing the probability of longer peace as well as a way to prevent multi-polar traps which commonly arise in game theoretic multiplayer scenarios with misaligned incentives (Armstrong et al, 2013 FHI technical report).

But what about the environment?

The interaction between individuals (or groups) and environments is the topic of Human Behavioral Ecology, and from that field we can sure gather that different dyads of a peoples and an environment will be conducive to different levels of altruism. Bulgaria and Turkey famously have inverse trust (Zak & Knack 2011), where individuals are willing to pay a cost to stop others from collaborating to a common resource pool that gets multiplied and redistributed. Whereas many northern societies are very collaborative, increasing a given pool of money every round in a fictional game of interest.

An altruist might want to maximize the quantity of population that lives in environments that are more conducive to happiness and to further compound cooperation.

Ethologists can also contribute to the happiness of animals by studying their own model and relation to the environment, especially those that humans can easily manipulate such as pets and livestock. Reducing animal stress or increasing their joy is a rarely practiced but impactful focus of attention for altruistically inclined behavioral scientists.

Seasonal altruism

The external milieu is not the only one that can change one's proclivity to altruism and cooperation. The notions of abundance and scarcity mindset, as modes of

thinking that activate different levels of altruism in our cognition, have gained some traction in popular science discourse. We evolved to think differently during harsher or milder seasons, and we can tap into that flexible cognitive toolkit to nudge ourselves towards a higher level of altruism. Increasing the perceived availability of resources, including food, land, mates etc… increases the probability of kicking our brain into abundance mindset, and far mode (versus near mode)(Hanson, unpublished). And in doing so, it increases the odds we will spend the system 2 cognitive manual time required to deliberate about the most effective altruistic decisions we could make to best benefit whoever it is we are trying to benefit. By contrast, if there is perceived scarcity (such as the toilet paper run during Covid-19) we become far less prone to allocate resources to others, and more stressed and defensive. In other words changing our perception, or our internal milieu (Keltner et al 2014), also influences our propensity towards altruistic actions.

Current political trends: Natural versus Sexual Selection

If we zoom further out even higher than seasons and food resources, we can observe a general pattern of periods in which the governing force organizing mating and politics is natural selection and other periods where it is sexual selection. During natural selection periods, scarcity creates the need for protective patriarchal prestige, gender relations are more polarized, right-wing systems and thought are more prevalent and women are relatively more submissive. During sexual selection periods, abundance enables more mate choice on the part of women who no longer depend as strongly on a man to sustain them and their babies. Riskier mating behavior and sexual power concentration happen, women and policies/parties favoured by women become more prominent. These periods roughly trace the oscillation patterns between the major parties in two party countries like the USA, and the left-right divide in multiparty countries like Brazil. Sexual selection periods generate fewer resources per male other things being equal, so they cause the relative scarcity that triggers the next natural selection period, with more resources produced, triggering the next sexual selection period. In both periods both Natural and Sexual selection continue to operate. Their relative strengths oscillate.

This is the Dream Time: For Sexual Selection

We live in a very anomalous time historically as our geometric rate reproduction, contra Malthus, has not yet caught up with the fast atoms technological development of

the 1860-1970 or the information tech bits development since then. We are unusually rich historically, as well as having those additional protective trust mechanisms. Given the combination of these factors we have been in a long sexual selection period, with increasing levels of feminism, both major parties moving left over a multi-decade span, and decreasing rates of marriage and children conceived in marriage. Further medicine put yet one more obstacle in the way of natural selection's filtering mechanisms. So our selection is currently strongly skewed sexual, and there was never a time in history where women had more political power than now.

Over time, it is possible that we fall back into Malthusian-like states and thus more patriarchal, male powered political systems. However, since technological change has made the fraction of the economy controlled by a small percent of people and their robots to be a majority stake, it is possible that we continue to spiral into more and more sexual selection as the strongest driving force in human evolution. Another possibility is that religious superorganisms, being the main groups that don't use contraceptives and out-reproducing secular groups, will tilt the political power back to masculine polarity without necessarily reaching the Malthusian barrier. We will discuss human superorganisms further ahead in more detail. For now suffice to say that life satisfaction for women seems to be inversely correlated with measures of gender equality (e.g. Women in Tunisia report substantially higher relative satisfaction than in Iceland), indicating that we are on a period of unusually low level of happiness for women given the lingering effects of second and third wave feminism. As of this writing, the direction ahead is unclear as feminism has lost popularity since peaking in 2015, but other indicators such as frequency usage of feminist specific jargon continue to rise in written and spoken media. Since gender relations are a red market (an oversaturated market with small marginal gains and a large quantity of participants) it is unlikely that an altruist would find sufficient leverage in an intervention that tries to reverse some of the welfare damage caused by feminism. The exception would be if there are small populations which are likely to outreproduce others such that there would be ripple effects. Reducing feminism among some highly fertile religious groups (Mormons, Hutterites, Amish) might be a leverage point worth considering.

Alienation as costly signal

Sexual dynamics create a further complication for altruism in that if you have concentric circles of power where the watchers on the wall are lower status than the protected noble families in the central castles, it may happen that political alienation, sheer lack of awareness, and also just alienation in general, become a costly signal of mate value. Take the clothing articles of 17th century Portugal and Spain for example.

Their exuberance is partially to signal that a nobleman who would wear it does not need to work to make a living. Similar for long nails and most of the feminine polarity adornments women use to this day in most countries. Costly indicators of being outside the grind and toil. In a sufficiently large society, the "ivory tower" and any other unusual set of beliefs discredited by most in a society, say, astrology in ours, can be taken as a signal of power in that to profess belief in it indicated you never had to deal with the consequences it would have if false.

Not knowing or understanding politics is a costly signal of power and mating value in a society that is very large. And it can be enhanced by professing the opposite belief of what those who deal with the actual borders would claim. To disagree with the margins is to assert one's mate-status as central. Consciously or not. So an altruist needs to balance access to the inner circles of power with a desire to know what is happening at the margins, so as to reallocate resources based on the ground reality, but without committing any faux pas that would detract their status as less than central to wealthy or high status peers.

Virtue signaling

Thus a balance is needed when attempting altruistic actions which takes in consideration that most people who attempt altruistic acts are virtue signaling (Miller G. 2007), to themselves and others, and need their actions to tickle the correct emotional centers to induce the behavior again. It is necessary to consider not only one's own tendency to virtue signal (for instance by allotting a fraction of resources for flamboyant signaling, so you spend the majority in actual altruism instead) but also the tendency of others. Creating prizes or premiums or other incentive hierarchies needs to consider that those who participate in such hierarchies have motivations of their own which include virtue signaling more often than not. Though some would reject this as immoral, it would, from an aggregative consequentialist perspective, be even more immoral not to allow virtue signaling in whichever quantity maximizes the actual good done. An altruistic agent should seek the Pareto optimal barrier where no change can enhance either virtue signaling or real impact without decreasing the other.

Immediate judgments of fairness and how they misbehave

Besides empathy discussed before, judgments of fairness can misbehave in a multiplicity of ways, many of which have been discovered by scientists.

Cucumber and Monkeys

In the most famous example by de Waal, some Monkeys go berserk if one is given a cucumber and another a grape for the same work, even if it is work they normally would do for a cucumber. Many humans reject unequal distributions as well.
We have multiple conflicting notions of fairness though. We split the bill in different ways in different cultures, not at all in others etc... Our intuitive emotional triggers for what is fair or unfair cannot be trusted, in that they often will cause an emotional reaction to A if B benefits us, but the reverse when A benefits us. That is, part of our felt sense of an emotional response regarding fairness is just selfishness masquerading as a moral outrage. Many human groups also have spirits and deities which play a role in judgements of fairness, either through their supposed opinions or through being hypothesized participants in a division and thus deserving a share themselves.

Christmas is Nuts!

Between ourselves we also have interesting ways to fairly distribute resources. Christmas triggers abundance mindset by having proxy fruits on a proxy fruit tree (Pageau 2019), and thanksgiving as well as Christmas involve taking concentrated sources of energy like nuts and meat and ingesting it in a way that increases probability of survival in the winter (in the northern hemisphere, where it evolved). We assemble together family and surrogate kin, and we assemble our selves at the moment and our selves at future moments, and we distribute resources in abundance, giving our future selves a full belly, and our family members with fewer possessions, food and presents often more costly than they could afford. Perfect to prepare for a harsh but predictable environment that selected for the K-selected mating patterns practiced by Christian human groups in the northern hemisphere. Kropotkin famously suggested that abundance itself was a cause of altruism at times and though much of the scientific community shunned his ideal as socialist pseudoscience, big game hunters and Christian families in December often demonstrate that his point was not devoid of merit.

Culture is fixed, biology is mutable

Contrary to much political hypothesizing, altering culture is often more resource costly to an effective altruist than altering biology. Most of the increase in IQ observed in the 20th century was caused by better sources of iodine and other nutrients in the food of the World's most needy. Although IQ is no longer increasing, and G has actually been decreasing since 1780 (De Menie, personal conversation). While educational

interventions seem to fail across the board except for tutors, altering biology with for instance ADHD medicine or higher levels of iodine has been systematically effective as a way to improve the quality of life, wealth and welfare of peoples. Because culture is more abstract it seems to our naïve eyes that it is also more flexible, but just like in human behavioral ecology there is often strong attractors organizing an equilibrium of environmental factors and a behavioral disposition, likewise culture often has strong attractors, which if we try to change just push back to their previous equelibria states. Some famous failed attempts at changing culture were trying to dodge the Westermarck effect by mating within the same Israeli Kibbutzim, Stalin going back on his attempt to destroy the family unit after realizing unrestrained female sexuality was causing even more problems than what he saw as the problem of family, and education among sub-saharan populations in many countries. An altruist attempting to maximize their personal impact might look for neglected interventions that change the biology of an area, and assume the cultural consequences occur downstream, and not vice versa. For a counterpoint on how culture also changes biology, see *Not By Genes Alone* (Richerson & Boyd 2005).

Pathogens

Pathogen prevalence is correlated with authoritarianism, and the separation mentality of high orderliness, high disgust sensitivity conservatism probably evolved to prevent pathogen infection (Fincher et al 2008). With the Covid pandemic, the world watched in real time as most countries, even the most liberal anti-conservative ones, closed their borders to travelers from one or other country, and sometimes to all foreigners.

Scarcity and high pathogen prevalence reduce trust, which decreases proclivity to altruism.

Although system 1 altruism levels decrease in a situation of scarcity and pathogen prevalence, system 2 can still conceptualize more complex points involving how other people might be faring even worse than oneself, which could increase coordination and generosity temporarily. During the initial stages of the Covid pandemic, dozens of the hundreds of scientists I know in social media were dedicating their time and effort to produce the most digested and relevant information as soon as possible, to help avert the worst case scenarios for the world as a whole. These individuals were not being rewarded for this behavior financially and the effect of that research in their own health is negligible compared to the time invested. In short, they were abstractly being altruistic towards the globe through system 2 deliberate effort.

Sanctity as the Altruist's tripwire

Sanctity is one of the moral foundations (Haidt & Graham 2007) that constitute how we think of valuable things. It has mate-guarding components, pathogen avoiding components, and superorganism cooperation components. When we hold something as sacred, it often is something that regulates mating, prevents disease and increases group coordination or de-coordinates with other groups. The LDS church has a "proclamation of the family" which is sacred, and the Black Lives Matter movement had an anti heteronormativity manifesto which, if you set them side by side, are almost inverted signs of one another. Both sides hold their document as somewhat sacred. Sacredness isn't a property of things in themselves but a relation between dyads. X is sacred to Y.

From an altruist standpoint, there is a constant danger that lurks in any form of sacred idea, space, or institution. To illustrate, pretend you are a white collar criminal. You can choose between two similar institutes to join and commit fraud and other crimes. The only distinction is that one of those is considered sacred by a substantial community in your city. Naturally, you should pick that institution so as to be protected by those who are devout to the sacredness of that institution. So an altruist should always be attentive to what a community holds as sacred value, because it is the best place to hide a destructive plan. The two most famous cases of something sacred being sequestered by antagonistic agents are the pedophilia scandals in the Catholic church and the political takeover of the SPLC away from its original desiderata.

The Fuel and Fire of Thinking: melting commonsense

Analogy as the lever of cognition and emotion

If we naturally use analogies and their subcomponent metaphors for cognition and emotion, if they are our levers for thinking, there should be a limit, a height above which we should no longer trust them to guide us correctly. In the same vein as sacredness or empathy have limitations as guiding lights for altruistic behavior, the general class of metaphors must also only go so far.

Breaking the lever

Analogies rely on narrative frameworks and descriptive scenarios. This creates a big problem in the case of scope sensitivity. In some experiments, individuals were willing to give more to help one person than 2, than 10, and than a 1000. Our cognition cannot conceive of large numbers and multiply benefits by them. We need to switch modes from the narrative, verbal, metaphorical cognitive mode for smaller scale altruism, to a mathematical, shut up and multiply approach when it comes to experiments of scale.

Burning the lever

We need in other words to burn the lever that brought us to the ability of conceptualizing actions as altruistic, ethically desirable, etc... and keep the end result, the flames through which we are measuring value. We need to quantify the good itself, or risk getting lost in scope insensitivity bias.

Quantifying good: QALYs

That's where QALYs and DALYs come in handy. These are measures often used in public health policy which allow us to quantify the quality adjusted life years provided by an intervention or the disability adjusted life years. They allow us to let go of "how we got to the good thing" and to quantify the good thing itself, so as not to get lost in large numbers, and still be able to compare the resource per good done ratio. The fewest resources and the largest good, the more effective the altruistic act.

A Moving Target: facing the void in effective altruism

That transition to quantifiable, alongside our previous directives to maximize, aggregate, and be global in time and space, constitute the central tenets of Effective Altruism: A movement dedicated to finding the most good one can do with one's life, and then doing it.

Or at least with part of one's life.

Effective Altruism is a group of people whose values are broadly in alignment, whose defining characteristic and goal is precisely to be more altruistic. Or, to be accurate, *most altruistic*. Yet, most instances of groups of altruists have a very different origin story. A deeper, older, and frequently sacred story. That is the story of the evolution of religious groups. We will now turn to that examination. The evolution of religion is in great part the evolution of a series of adaptations that are beneficial at the group level, and those often necessitate the collaboration or acts of altruism between the individuals of which the religious group is constituted. Sometimes the religion is simply a group that often favors its members above other members. Other times however, under a more complex evolutionary lens we will observe that some religious groups are full fledged superorganisms.

Tools of The Trade: Superorganisms and Religion

"Before we rejoice at the death throes of the relatively benign Christian religion, let's not forget Hilaire Belloc's menacing rhyme:
'Always keep a-hold of nurse
For fear of finding something worse.' "

— Richard Dawkins

Let us try to analyze religions and to investigate to what extent they can be seen as biocultural superorganisms, which are tightly knit systems which necessitate constant intraparties altruism.

My hope is that evolutionary anthropology provides a unique perspective through which to read the biocultural unfolding of religious groups and other human social units, and that the understanding coming from areas as diverse as philosophy of biology and psychology can bring us to a secular understanding of some aspects of branches of Abrahamic, and in particular Christian religions that have remained unexamined throughout much of their history, and which can be extrapolated to some other religious groups, some nations and some ethnic groups.

My argument has as prerequisite the work of many researchers from different fields, and it may be easier to understand it by explicitly mentioning these influences:

From cultural evolution researchers Boyd, Richerson, Lewens, and Blackmore, I bring the biocultural evolutionary framework.

From Herbert Spencer, E. O. Wilson, and frequency-dependent evolutionary biology, I bring the notion of a superorganism to bear the similarity between superorganisms in eusocial species and religious groups as superorganisms.

From biologist David Sloan Wilson, I bring the lens through which to read evolutionary processes as being exerted on units of selection. These units can be groups, individuals, genes or entities of a different biological resolution under different selective pressures.

From philosopher Ellen Clarke, we bring a technical framework through which to conceive of when groups become individuals, and what individuality is from an evolutionary perspective that is substrate independent, and thus applicable to biocultural, cultural and biological entities alike.

From neuroscientist and anthropologist Terrence Deacon, we bring in a physicalist conception of telos and directionality exhibited by biological systems and processes.

From evolutionary anthropologists McElreath and Joseph Henrich, we bring the assumption that different religious communities have evolved patterns of behavior and beliefs that have adaptive value.

From personality and religious psychology professor Jordan Peterson, we bring a reading of the psychological value of religious narratives in the bible conceptualizing them as evolved narratives that capture fundamental aspects of personality transformation, with an emphasis on the individualization of the self and the sacrality of the individual.

Our investigation aims to examine the evidence for conceiving of religions as biocultural units of selection in competition with other units such as genes and different sized groups, to examine the consequences of a specific adaptation most easily observed in Christianity, the alliance of the superorganismic level of evolution with the individual level.

Human groups as superorganisms

Why the notion of human groups as superorganisms keeps emerging

The notion of human superorganisms is intuitively appealing; one just has to observe for a while the interaction of a socially tight group of humans to find sophisticated displays of both synchronized behavior and self-restraint by individuals which appear to favor the group they're a part of. Such behavior can be seen for instance in religious communities, sports teams and military units

Our very language is ridden with expressions suggestive of the ordinary conceptualization of human groups as individual agents on their own (fear the "mob rule"!). We have amassed a large vocabulary of *singular* nouns to refer to several types of human assemblages - such as 'gang', 'clan', 'tribe' and 'horde'. This framing of aggregates of human beings as individuals has become entrenched social reality across several institutions and academic disciplines. For instance, in military hierarchy, groups of men are assorted in several differing 'units' of functional organization (such as companies, platoons, squads, etc). One of the most influential works in management and organization studies has human companies and organizations being metaphorically cast as both living organisms and gigantic brains in order to elicit greater understanding of how they work (Morgan, 1986).

Prima facie, this way to look at human social evolution seems to clash with the "selfishness" of the gene-centered view under which much of evolutionary thought is predicated, thus begging for an explanation. According to Richard Dawkins (2017) evolution is said to be selfish and myopic, but when groups compete against other groups, collective selfishness and longer term goals can supersede those of the individuals involved. Much like the increase in volume of animals from unicelular to

pluricelular enabled and stabilized somatic adaptations that are self-undermining in the short term but that provide evolutionary benefit in the long term, so did the emergence of group as a unit of selection, in humans (especially bio-culturally tight groups) enabled the stability of adaptations that are self-undermining even in the long-term — for an individual — as well as strategies that are only stable at populations levels.

The notion of groups as units of selection is frequently criticized, especially among cultural anthropologists, but also in the social sciences at large. Thus, for example, the work of Herbert Spencer (1864), who coined the term 'superorganism', is frequently conceived of as being a dangerous precedent for reasoning that motivated undesirable political and social movements throughout the 20th century.

However, its value lies in providing tools for examination: The hypothesis that human groups are analogous to superorganisms allows for rich exchange of ideas between those domains, providing conceptual tools for anthropologists and entomologists alike to attempt a better understanding of natural phenomena that are often hard to explain with a simple evolutionary lens. A decade ago, the most important compendium on superorganisms, *The Superorganism* (2008 Holldober & Wilson), provided the canon on thinking about that concept in insect eusocial species, which renewed interest in other domains in which it could apply. Concomitantly, a new proposal to formalize evolutionary mechanisms of group adaptation has given a mathematically precise description of superorganisms (Gardner & Grafen, 2008). Ever since, novel theoretical developments of the superorganism account have been devised for human social groups both in general (Kesebir, 2012, Aunger, 2017) and in particular (Duarte et al, 2012, in the context of sports teams).

Defining Superorganisms

Here I recognize a group as a superorganism if it satisfies the following desiderata:

(1) it can be parsimoniously described as a group of agents;

(2) these agents are individually distinguishable;

(3) evolutionary pressure is exerted at the group level to a sufficient extent to make the group a unit of selection (Wilson & Wilson 2007);

(4) there are frequency-dependant roles or functions exerted by those agents that are codified, genetically or symbolically;

(5) the group is capable of some responsivity to external interference on its *telos* (Deacon 2011) in part in virtue of the existence of these different roles.

(6) a collection of single creatures that together possess the functional organization implicit in the formal definition of an organism.

Here (1) Distinguishes superorganisms from organisms. (2) Distinguishes superorganisms from individuals with organs. (3) Delineates the framework in which our analysis is done, the evolutionary analysis commonly utilized in the literature on biological evolutionary transitions and multi-level selection. (4) Distinguishes superorganisms from units of selection simpliciter. (5) Distinguishes superorganisms from group selected agroupments without agency at the group level. (6) Distinguishes it from other systems not composed of unitary individuals functionally organized as an organism.

Considerations of agency are crucial to characterize superorganisms. The superorganism is a "purposeful being" with an "agenda" or goal content structure or telos of its own (Gardner & Grafen, 2009). By theoretical default, the behavior of the superorganism aims at the maximization of the biological reproductive success of the group. However, in a symbolic species such as *Homo Sapiens* (Deacon 1997), the evolutionary dynamics (Nowak 2006) can get much more complicated due to a concomitant objective function pursued by cultural replicators co-evolving with its primate hosts (e.g. Henrich 2016) - objectives which may or may not converge (Blackmore, 2009) as well as form partial coalitions and other game theoretic structures.

The behavior of ant colonies, including their complex architectural extended phenotypes, displays extraordinary "competence without comprehension" (Dennett, 2017); ants only have "free-floating reasons" for their behavior (Dennett, 1995).

Members of symbolic and cultural species may represent and reflect upon their reasons, even about those we have no conscious access to, such as the cognitive unconscious, self-deception (Trivers 2013) or *The Elephant in The Brain* (Hanson & Simler 2016). The collective agency explicitly exhibited by a human social group may also be of a distinct character (at least in part); they are a function of *joint intentions*, predicated upon information about shared goals and recipes for coordinate action being publicly available through language (Bratman 1993, Partenotte 2016).

As delineated by Hamilton, Smith and Haber (2009), under the approach to individuate superorganisms by appealing to evolutionary trends of selection, there exists theoretical disagreement. From the Wilsonian perspective, which emphasizes a more continuous transition from colony (a population of conspecifics living in proximity) to superorganism, becoming a superorganism is a matter of between-group competition overcoming within-group competition. Under the approach of Reeve and Hölldobler (2007), within-group competition needs to be nearly nonexistent. In Ellen Clarke (2013) model, the transition between group and individual is given formal philosophical definiens.

If my hypothesis is correct, the coalescence of human religious groups with their respective religious texts, customs, norms, and rituals forms an evolutionary structure with its own telos that partakes in both intra-level competition - with other

superorganism heavily constituted by human groups, such as some nations and some religions - as well as inter-level competition, with its constituent individuals and with supra-religious categories which may be nations, cultures or other abstracta that partake in evolutionary competition over time.

Emergence of human superorganisms

"Our genes, it turns out, can eat their cake and share it too."

— Joseph Bulbulia and Marcus Frean

How could a population of encultured social primates transition into a fully-fledged superorganism?

Intergroup selection led diverse groups of humans to act on behalf of the group. Boyd and Richerson (2009) mention that reciprocity and reputation can explain the stability, though not the emergence of large scale cooperation. This form of intra-group cooperation, even when large scale, is called parochial as long as it is not indiscriminate. That is, it is cooperation that favors or biases in favor of a specific set of individuals that contains oneself. The emergence of cooperation might have occurred via a process of social addiction, in which our species self-domesticated, becoming as it were addicted to the presence of others, and losing individual capabilities, which were offloaded presumably from individual cognition to group intelligence.

Group intelligence above and beyond individual intelligence is a hallmark feature of the superorganisms in the insect world, where improved cognitive capacity at the level of groups supersedes that of individuals in a variety of decision-making contexts, such as perceptual discrimination (Saaki & Pratt, 2018). There exists a sizable literature scrutinizing the contexts and mechanisms under which human groups may collectively outperform the cognitive capabilities of individuals (Kerr et al 2004, Kurvers et al. 2015). Also, the study of synergistic effects of human beings coupled with digital communication devices has been revolutionizing solutions to complex social problems such as the prediction of elections, company management and product development (Krause et al, 2010). In the field of human-assisted computation, striking parallels between the patterns of social dynamics of internet-based social networks, such as leadership maintenance and the bifurcation of existing projects, with the collective behaviors of colonies of eusocial insects have been delineated (Pavlik & Pratt, 2013).

Cultural transmission increased group-level heritability, as cultural information is frequently likely to be transmitted within but not outside the group, creating new selective forces that put the unit of selection of groups potentially ever closer to the forefront of the evolutionary process (Soltis et al. 1995, Bloom 1997). As this new unit of

selection began to be more relevant as the one upon which evolution was acting, group-friendly optimization processes became stronger leading to snowball effects in cases of cultural solutions to coordination problems - e.g. The commandment to respect Sabbath, usually coordinating observant Jews on Saturdays and Christians on Sundays.

Further, different societies, with different evolutionary pressures, adapted distinct levels and types of intragroup prosociality and intergroup conflict resolution mechanisms, often proportionately different to their biogeographical conditions (Diamond 1998, Sowell 2016, Spencer 1873), where local conditions favor different interpersonal behavior. Some of these locales favored more prosocial punishment - such as punishing defectors - others, antisocial punishment - punishing the most altruistic participants (Sapolsky 2017) - the reasons for that are not yet well understood. Joe Henrich (2016) mentions that larger societies tend to have larger Gods as well, Gods who are believed to have a wider scope of action. Monotheisms have scaled substantially more than non-monotheisms in areas that are less isolated and multiethnic.

Many distinctive mechanisms to reduce intragroup conflict become available through culture, from the cultural modulation of prosocial emotions meant to elicit group cohesion to the establishment of complex egalitarian institutional practices such as the rule of law (Kasebir, 2012). Mythology, storytelling, music and even punishing Gods (Henrich et al 2016) facilitate this process.

When the loss of autonomy happening as a result of cooperation and offloading individual intelligence to collective intelligence is sufficiently strong, the 6 threshold conditions above might be crossed, thus creating a superorganism. This process occurred several times in biological evolution, and arguably occured in biocultural evolution among some human groups.

Many human structures approach these thresholds. Human cities, whose life expectancy outlives that of religions, nations and dynasties, are progressing towards a similar relation with the environment and change over time. As a biological unit, they are far too permeable to count as superorganisms - they don't have enough attractor stability and fidelity over time to be discretized in a way such as to permit a full evolutionary analysis - but as a biocultural unit, their boundaries and borders might be "thick" enough to make them susceptible to selective and adaptation forces. Any sufficiently well isolated group of biocultural entities can become a unit of selection if the adaptive forces involved coerce it the right ways. It is possible to see the emergence of superorganisms in purely biological strata, in purely cultural strata: with intentional memetic systems (Aunger 2000) such as corporations or even music being the organisms that evolve together, and, more interestingly to our discussion: in biocultural evolution, that is, evolution involving both biology and culture, as is the case in many political units, religions, dynasties, and castes.

In the most recent years, *Darwin's Cathedral* by David Sloan Wilson was the main thrust in the direction of considering cultural groups such as religions as a unit of selection on the forefront of biocultural evolution.

Once the prospect of group selection is taken as part and parcel of evolution, more subtle questions begin to emerge. Questions such as: is there a way to determine which level of complexity is better suited to describe where the fitness pressure lies in a system? This sometimes can be determined through the use of the Price Equation. Or questions such as can Multi-level selection imply multiple levels are concomitantly partaking in the selection process? It doesn't prevent that, but also does not necessitate it. And most interesting for our purposes, it can mean multiple levels compete against each other, and alliances can form between different levels of organizational complexity. That is, inter level competition can lead to between level coalition. I call this intralevel, multilevel and interlevel complex by the umbrella term 'biocultural' evolution in this text.

A process that is relevant to understand competition between levels of selection - intra-level competition - is not unique to superorganisms, our cells and organs frequently take actions that can be described as competitive actions against the organism as a whole, which the organism has regulatory mechanisms to prevent from going too far astray from homeostatic resource usage equilibrium.

As E.O. Wilson & David Sloan Wilson (2007) sum up, "Selfishness beats altruism within groups. Altruistic groups beat selfish groups. Everything else is commentary."

This view is not unique to biology. Anthropologist of cybernetics Andrew Pickering describes an ontology proper of the British cyberneticians in which dances of agency constitute the primal ontological forces operating in systems that reciprocate with one another. This perspective, as I see it, can gain much traction if aligned with Eva Jablonka's model of 4 dimensional evolution where the entities that partake in evolutionary competition become substrate independent, as they are composed by a combination of ever evolving, changing systems of four different kinds, in what can be thought of as a dance of attractors, of telos, of levels of evolutionary selection constantly subject to forces which change their morphology while at the same time, in the ways we will see are described by Ellen Clarke, making them individual or group along the way. To Pickering, this dance of agency of cybernetic systems has a main feature: no dance partner in the dance is the main agent.

Transitioning to Eusociality and Superorganism: The Risks

Perhaps the most profound difference between prototypical superorganisms - such as colonies of eusocial insects - and human groups is the degree of genetic relatedness of its constitutive individual organisms. By endorsing "low-levels of heritable within-group variation" as a defining feature of superorganism, Selin Kesebir (2012)

argues that this condition can still be aptly fulfilled in human superorganisms even in the absence of the genetic homogeneity that we observe in eusocial bees or ants. This stems from the hypothesis that most phenotypical variation in humans is largely cultural in origin, with cultural and geographic evolution operating sufficiently similarly to biological evolution plus a wider view of heritability that encompasses geographical and cultural forms of vertical transmission of information, from parents to children. Phenotypic similarity across human groups could then be enforced through the social acquisition of norms (e,g, Amish head cover, YouTube makeup tutorials, suits, hairstyle, etc…) through learning, contagion and co-location.

Here, the account of human groups as superorganisms finds synergy with the research on the evolutionary basis of ethnic nepotism (Salter 2002, Salter & Harpending 2013). This broad research program sports many theoretically attractive features, such as the enabling of potential group-selectionist mechanisms. Through culture, an existing natural tendency of altruistic behavior towards co-ethnics can get amplified alongside the social monitoring mechanisms that allow the identification of free riders and other defector individuals as well as their consistent exclusion from the group gene pool (Campbell, 1983). The mechanisms underlying ethnic nepotism can simultaneously contribute to the explanation of positive assortative mating and the formation of castes or subgroups inside a social group.

The Risk of Distributed Intelligence

One risk underlying higher levels of selection is that of intelligence itself becoming distributed among the agents that constitute an evolutionary unit. Natural selection isn't particular about installing intelligence inside individuals or as a network effect of the combined actions of multiple organisms. This is a problem because we don't have access mentally to at least some valuable data or levels of abstraction that might be useful in facilitating altruism. This is analogous to the economist's complaint that communist regimes could not assess and process pricing well with a centralized economy, leading to scarcity or overpricing. Likewise, our understanding of things like existential risk (Bostrom 2013) is incredibly tenuous and in the minds of very few, very smart people. So although higher levels of selection can help myopic evolutionary success, it might betray us in more complex survival tasks, such as surviving the next millennium. To counter these forces, religions and States that push towards individualism, such as Christinaity and the USA might end up being the literal difference between life and death to our species and all the things it could eventually produce of moral value.

Clines and ethnic groups as superorganisms

Over the last decade, advances in bioinformatics have pushed into great heights research in anthropological genetics uncovering the genetic substructure of human populations (Tian et al 2008, 1000 Genomes Project Consortium 2015). Clusters and nested hierarchies of genetic similarity which approximate the referents of many prevailing sociopolitical terms (such as ethnicities and nationalities) can be readily extractable from genetic data.

Against a prevailing consensus which entered the public sphere in the seminal paper published by evolutionary geneticist Richard Lewontin (1972), the late anthropologist Harry Harpending (2002) has estimated the kinship similarity between two random individuals from an arbitrary human population to be 12.5% - equivalent to the kinship coefficient between uncle/aunts and nephews/nieces. This was before the development of current DNA microarray technology and employing an older data set. More recent estimates using contemporary genomic data from national populations have vindicated Harpending's early predictions of substantial kinship within human populations (Salter & Harpending, 2013).

Ethnic groups are thus an obvious candidate for superorganisms due to the shared genetic similarity of their individual units through relations of kinship. This is a view with some historical precedence. Herbert Spencer's *Sociology* (1873), for instance, already conceived of the evolution of human groups from this perspective. More recently, Howard Bloom's *The Lucifer Principle* (1995) also gained popular attention by presenting the thesis of human groups as "social superorganisms".

This perspective can be mathematically useful for modeling population dynamics. However, using ethnic groups as the sum total constituents of superorganisms has also been argued to be unwarranted in a few ways:

First argument: the ethnic view demotes the importance of culture, giving prominence to biological features that are more salient but not necessarily as important in determining the flow of the evolutionary process, or the determination of boundaries for unit of selection evolution.

This argument has a counter-argument, however: In birds and frogs, the first aspect of separation of a species into two is usually phenotypically visible. A birdsong begins to differentiate into two, and the females select one or the other song, leading to progressively increasing differences and, if continued over a sufficiently long timespan, speciation (Farias-Virgens, M., & White, S. A 2017). Visible differences that make a difference for sexual selection are good indicators of the beginnings of formation of a sexual chiasm. Most people in our globalized world predominantly mate with people of the same race and until recently also from the same small sub-population. Showing some degree of separation on the outside does substantially increase the probability of following separate evolutionary paths, much like races of dogs or other breed animals.

A second argument concerns the demotion of culture: I argue that culture is not a free-floating variable and that it should be contextualized in light of any of its possible biological underpinnings - such as, for instance, differences in average group genetics

that influence personality. As Boyd & Richerson (2009) put it, “culture is constrained by social instincts". For a specific interplay of biology with culture, differences in the cultural value axis of individualism-collectivism have been found to correlate with differences in the allelic frequency of variants of the serotonin transporter (Chiao et al, 2010) and the oxytocin receptor (Luo & Han, 2014), which are basic neurobiological underpinnings of social behavior and cognition. The socioendocrynology of different peoples (Ellison 2009) should no less be discarded than age differences or sex differences in socioendocrine modulated behavior. A consilient view of biology and culture, elucidating its complex causal pathways and patterns of interaction (Wilson, 1998) makes for more empirically adequate and responsible anthropology.

Third argument: the ethnic view is usually considered dangerous as putting the fact and value entanglement that took place in the beginning of the 20th century at risk of taking place again. In 2009 social psychologist Johnathan Haidt declared that the “most offensive idea in all of science” is “the possibility that behavioral differences between racial and ethnic groups have some genetic basis”. Whether or not academics consider biological differences between ethnicities problematic, as the information percolates into the general public, it becomes political and moral. Caution is necessary to prevent it from being utilized as a justification for oppressive or otherwise strenuous relations between different groups or political clusters. This view is countered by another line of thinking - e.g.Pinker (2004) - suggesting that precisely the hiding of relevant information about group differences, whether between sexes, races, or otherwise can lead to disparities being confused with discrimination (Sowell 2018) and thereby hiding potential avenues for improvement of those different groups in different axis.

Concerning the moral turmoil of the scientific investigation of human biological diversity, we must be wary of the moralistic fallacy, the inference from how things ought to be to how they really are, if science is to seek truth. But as someone who is deeply concerned about social oppression and the reduction of global suffering, I argue that factual truth about the etiology of social oppression and inequality are necessary for the design and implementation of social policy that is more effective at the achievement of peace and social progress. This has been argued by many scientists and philosophers (Singer 1979, Pinker 2004, Anomaly 2017, Winegard et al. 2017, Sowell 2018).

Fourth argument: the ethnic view doesn’t seem to capture the right level of resolution for the evolutionary forces to be acting on. Often it is argued that the individual is the level at which evolution takes place. Other times, smaller groups (a single tribe for instance) are taken to be a more significant unit of selection than a cluster determined based on phenotypic invariances that are apparent to the human eye, such as our determination of races.

For a potential case study of how the view of ethnicities as superorganisms can be fruitful, consider Haidt’s (2009) conjecture of how we may uncover group differences in the trait of “clannishness”. Here is an example where it is possible to investigate a natural

propensity towards ethnocentrism by uniting the superorganism account of human ethnoreligious groups with evolved ethnic nepotism. Human demes differ on their levels of genetic relatedness (which can be measured, for instance, through inbreeding and kinship coefficients) and it is an open question whether the reduction in within-group genetic variation in human demes could be significant to enable superorganismic cohesion. It could also be a partial but not sufficient condition for superoganismic levels of organization.

Religious groups as superorganisms

Religion is a very important aspect of most of humanity's lives on a global scale. Aggregating the World Values Survey's seven waves of data, respondents from over 100 high- and low-income countries, 48.2% said they prayed everyday, and on a 1-10 scale, the median importance respondents asserted that God has in their lives was 10. Moreover, the *societal* salience of religious beliefs was clear: 69% of married participants share their religious beliefs with their spouse; 20.2% of respondents mentioned members of another religion as a group they would not like to have as neighbors; and 79% considered religiosity a "very important" trait in a woman (Inglehart et al. 2014). It is noteworthy that marital preference is likely to express itself into a higher heritability coefficient of religious affiliation through mate sorting, self-segregation and segregation.

Evidently, as opposed to being merely a matter of private individual belief, the way New Atheists such as Dawkins (2011), Harris (2006) and others characterized it, religion is very much a social endeavor.

Indeed, religious groups often show higher intra-member genetic similarity than would be expected by chance (Haber et al. 2013; Ostrer 2001), a pattern that emerges both top-down from the direct influence of religion in its members' marriage patterns, and bottom-up due to the genetic influence on psychological factors that affect religious affiliation (Bradshaw 2008) and the mere geospatial propinquity between members of a religious group that fosters endogamy - although the latter factor has been alloyed by the growing globalization and the resulting large-scale European conversion of global native peoples.

Religious groups also show some degree of cultural homogeneity — religious orientation is one of the most important factors influencing people's moral and metaphysical beliefs, as well as their everyday habits and ritualistic activities, and some argue it is precisely such collective synchronicity and moral consensus that religion brings that contains its value for humanity and the reason why most people are religious (Graham and Haidt, 2010; Boyer, 2001; Durkheim, 1915/ 1965).

David Sloan Wilson, in his 2002 book Darwin's Cathedral, takes such observations of the inherently social character of religion and its power to bind groups together

genetically, culturally and behaviorally to their logical conclusion, and analyzes religious groups from an evolutionary multilevel selection perspective, arguing that religions can be seen as units of evolutionary selection. Such a claim is not quite as audacious as that of religions having superorganismic qualities, as it only requires that they, as units of selection, compete with other units such as genes and individuals, but it points in a rather similar direction. Adding functional role differentiation, and approximate telos to that, and the superorganismic view naturally emerges.

Most importantly, Wilson opened the field for other discussions on the evolutionary significance of religion, bringing the academic spotlight to seeing its essence as lying not in the belief in God *per se*, but in "the relation between man and his fellows," as he quoted from Isaac Bashevis Singer in his book's epigraph.

Wilson argues that religions have *secular utility*, highlighting evidence of the increased rates of cooperation between members of one religious group compared to the rates between members of different ones with reliance on costly signals, both from laboratory economic studies and historical evidence (Iannaccone, 1992, 1994; Henrich, 2009; Bulbulia, 2011) and emphasizing the encouragement most religions give for its members to reproduce (as in "Be fruitful, and multiply," from Genesis 1:28).

Almost tautologically, reproduction is important for evolutionary stability, and one of the most prominent features of successful present-day religions is their encouragement of reproduction of its members. Indeed, in the last centuries of the western Roman empire, the cradle of Christianity, paganism and other features of the majority's social structure highly discouraged reproduction, and the population was in decline. Christianity preached the precise opposite of that — just like Judaism, from which it originated, it expected marriage and children. Today, the Roman Empire has long collapsed and paganism is nearly unheard of, but Christianity lives on as the largest religion in the world, and to this day, religious people have higher marriage and fertility rates than non-religious people (Pew Research Center, 2014; Inglehart et al. 2014). Some groups, notably Mormons, have fertility rates higher than replacement and missionaries who convert distant non-Mormons to their religion in addition.

Sociologist Kevin McQuillan (2004) has delineated three conditions which if satisfied enhance the prospects of religions to achieve demographic influence. These are the prescription of behavioral norms which reliably cause fertility, the enforcement of social compliance to these norms and the existence of a strong sense of attachment to the religious group, "raising the rewards for compliance and the penalties for deviance" (2004). Many of the members of the cluster of mores commonly associated with traditional or conservative religious morality - such as the promotion of heteronormative marriage, the opposition to sex acts outside vaginal intercourse, the demotion of contraception and the condemnation of abortion - which could at first sight appear to be irrational and idiosyncratic, can be readily interpreted as strategies that increase fertility. Measures of religiosity have been positively correlated with fertility intentions and actual fertility (Frejka & Westoff 2006, Hayford & Morgan, 2009). In some cases this can be

elusive though. A superorganism distinguishes itself from a mere unit of selection in part by functional specialization of castes and groups. This can be seen in very clear form in the case of Judaism. Whilst conservative Jews tend to bear many children and partake in more observant and rigorous rituals, liberal non-religious Jews sometimes fulfill non-reproductive roles that may still favor the 4 dimensional structure, in Eva Jablonka's evolutionary terminology, of the Jewish superorganism, either by facilitating the survival and reproduction of the more reproductive caste, or by deteriorating the reproductive capability of competitors, a contentious argument most extensively developed by MacDonald (2002) which recently regained attention when it became an academic debate with Nathan Cofnas.

Despite those points, few academics have heretofore explicitly examined religion as a superorganism, a united organism with its own evolutionary fitness. Joseph Bulbulia and Marcus Frean explicitly examined religion as a superorganism in response to *Darwin's Cathedral*. Shade Shutters from Arizona State University makes a brief case for the full-fledged superorganism claim (Shudders, 2013), delineating some coarse-grained analogous structures between those two systems, such as the presence of behavioral policing and of different functional roles embedded in different agents, equating those behavioral characteristics as resembling insects in superorganisms.

Religiosity, Ideology, Ethnicity and the God Shaped Void

In Western Europe, identification with Christianity has been associated with anti-immigrant and nationalistic attitudes, implicitly denoting an underlying ethnic character (Storm, 2011).

Interestingly, an abandonment or religious attitude also correlates with an increase in nationalist attitudes. The implication seems to be that when people drop religion, they fill in what some call a god-shaped void with intragroup identitarianism, or with some all explaining ideology of a different nature. Those who are more left leaning drift towards the identity politics/postmodernism/Marxism/feminism attractor, those who are more right leaning drift towards either nationalism (e.g. in Brazil, Bolsonaro's election. In US, Trump, UK, Farage) or ethnic ingroup preference (e.g. Richard Spencer, Jared Taylor and different varieties of white identitarianism, Zionism in the case of Israeli Jews). This curious phenomenon of what is referred to as a God-shaped hole has persuaded thinkers as diverse as Nietzsche, Steve Bannon, Jordan Peterson and Dennis Prager. I believe understanding the psychological nature of this phenomenon to be one of the most important tasks for psychology, neuroscience, and socioendocrynology in this century.

A speculative hypothesis I have is that the framework one uses to interpret incoming information depends on an encoding pattern in the hippocampus that is expressed differently in different people, and that it switches between two discrete

encoding systems when a traumatic event switches an individual's political beliefs or metaphysical beliefs for instance in PTSD. This frame switch encoding transition seems to be responsible for the process of remapping the environment when a new danger or a new geography emerges (*Maps of Meaning* lectures, Peterson 2017) and I suspect that religion, ideology, and political orientation have a distinctive relata in these encoding patterns shown in Peterson's analysis of trauma in rats and humans and personal experience of transitioning between belief systems. Besides, it seems from De Dreu's papers on moral psychology and Sapolsky's inferences in *Behave* (2017) that our in-group out-group sense is modulated by oxytocin, which would explain the univocal nature of the *us them* encoding distinction represented by the encoding mechanism that was conceptually captured in the idea of a God shaped void.

Besides being potential constituents of superorganisms, this neurological conjecture would give us a psychological reason to believe religious groups to be similar to ethnic groups: Both are a potential filler to the God Shaped void in our psychology. This would give an explanation for the problem that to many people, only one psychological construct can occupy this God shaped void, either a religion, an ethnopreference, a commitment to the politics/postmodernism/Marxism/feminism attractor or some other construct, what I call *the ideological binding unicity problem*.

These entities, prima facie, seem distinct. I wager that the unifying factor they have is that they determine the encoding of what our brain, via oxytocin, distinguishes as an "us" or a "them." This explains the discreteness. If there is only one differentiation boundary us/them, then there can only be one void. This is the most parsimonious explanation I'm aware of that is compatible with moral psychology, socioendocrinology, behavioral genomics, the historical record, and the statistical patterns observed in social psychology, as well as my own experience and observation of people who entered or left an ideology, religious system or ethnic preference group.

Here we begin to see, if I am correct, that from an evolutionary perspective, to distinguish the categories of ethnicity from religion may be a category mistake. Sometimes either one of those, or a combination thereof, is operating at the superorganismic level.

A more game theoretic and price equation based way of noticing the similarity between these prima facie disparate categories is that both are a manifestation of a higher level of selection clamping down on a lower one.

Establishment of group boundaries and borders

Schelling Fences or "Why are cells borders where they are?"

In economics, theory of the firm contends that a firm will grow up to a point where transacting with outside entities is more profitable than internal transactions. Since evolution is an optimizing process, we should expect something similar to determine cell sizes. A cell should, in expectation, be as big as it maximizes the combined efficiency of its internal and external transactions. That is the Schelling point to put on fences that determine what is in and what is out. Systems with boundaries that evolve by local gradient ascent will usually find local Schelling fences. Stochastic systems (e.g. Hox duplication) may also find global schelling fences.

Dynamic equilibria are also possible but less likely, it may also be the case that evolution happens too fast through a Red Queen like phenomenon, leaving no stable global optimum, causing ideal size to change over time, following the gradient ascent of evolutionary search through phenotype space.

Why are organisms boundaries where they are?

For multicellular organisms, the number of considerations for expected size grows sufficiently much that simplified economic considerations like the ones above are insufficient to assess final size. Theory of the firm, gradient ascent and stochastic search are useful concepts to determine Schelling fences in the abstract, but in the biological world of multicellularity, complexity rules sometimes subdue the apparent conclusion of abstract extrapolations of these lower level phenomena into larger and larger substrata, for an extensive overview of similar processes, see *Incomplete Nature* (Deacon 2011).

Why are superorganism boundaries where they are?

Since superorganisms in the present analysis are composed of four dimensions (genetic, epigenetic, niche construction and cultural) as an evolutionary structure, they are better conceptualized as having gradient boundaries in some dimensions and better determined boundaries in others. This may seem initially damning to the concept, but as *Five Misunderstandings about Cultural Evolution* (Henrich, Boyd and Peterson 2005) explains in detail, the existence of attractors in concept space suffices for the preconditions for evolution to take place, and the same holds in the case of superorganisms. Although the boundaries of superoganisms are not very distinct in some dimensions, as long as there are attractors and constraints generating the conditions of production for their own continuity or re-creation (Deacon 2011) they can be seen as teleologically oriented and organismic.

Defenses biological superorganisms have for invasion

For Sober & Wilson (1988), the conceptualization of eusocial insect colonies as superorganisms allows for the metaphorical aptness of comparisons going from parts of prototypical organisms to the units of functional organization of superorganisms. A standard case would be conceptualizing soldier ants as part of the corresponding "immune system" of the colony. Many such analogies can be delivered for human societies conceived as superorganisms; we have systems of transportation enacting the role of distribution of matter and energy, much like circulatory systems, and refineries and processing plants acting like its converters, much like the digestive and respiratory systems (Heylighen, 2007). Superorganisms have some forms of defense against invaders which admit of analogies in human groups. Soldier ants were conceptualized as soldiers because like human soldiers they are a separate caste who dresses up in different phenotypic expression and perform specific actions that may not be personally conducive to maximal fitness, but still might lead your group to be more fit than competitor groups, and thus epigenetic combinations that can express soldier like behavior remain in the gene pool of humans and ants alike. Evidently among humans much of the evolution of soldiers is culturally determined and not purely biological/biochemical, as seems to be the case with ants.

Not all superorganismic defenses involve overt physical aggression (or threats thereof); the collective notional worlds of groups of humans have also evolved a myriad of defense mechanisms against *cultural* invasions. For instance, in Christianity, the characterization of ideological dissent or opposition to established doctrine as "Satanic" is a defense mechanism that allows the preservation of social identity. In the theory of identity-protective cognition (Kahan, 2007), challenges to collective beliefs can be interpreted as attacks directed at the group. Likewise, the first four commandments in the Hebrew Bible clearly distinguish game theoretic collaborators from non-participants in the Judeo-Christian behavioral game, thus serving as strong and hard to fake markers of a participant in a group and a set of behavioral predispositions associated to them.

Cultural groups and their defenses/boundaries

Cultural groups can have several different boundaries that constrain their evolution, and sometimes also organize (literally as in: creates different organs) their internal structure. We've already considered this in the context of religion.

Religious Nationalism

"We may be in the throes of the discovery that the only thing worse than religion, is it's absence."

- Douglas Murray

Religious nationalism can be a species of either ethnic nationalism or of cultural nationalism. As already pointed out, religions can be analyzed from the point of view of an axis of ethnocentrism and universalism. Ethnic religions reliably produce ethnoreligious groups through assortative mating.

Contemporary religious nationalist movements confront secular varieties of nationalism which do not place religion in prominence as the most fundamental basis of social identity (Juergensmeyer, 1993). This religious grounding serves to establish international solidarity and agonism. For instance, a strong allegiance by far-right Russian nationalists to the Eastern Orthodox Church serves to create antagonism and differentiation from the Western world which is largely of Catholic and Protestant extract (Verkhovsky, 2002). Fully fledged religious nationalism, stripped of ethnic commitments, can be hard to come by. For instance, the religiously observant (non-secular) varieties of Zionism are inseparable from Jewish ethnicities. As of Oct 2017, even mitochondrial analysis became grounds for Israeli citizenship, making it the only *officially* ethnic state. The nationalist ideology of Hindutva in modern day India is commonly framed as a religious movement but the way Hindus are conceptualized is religiously inclusive, encompassing adherents of traditional religions in India including Muslims, Jaina, Buddhists and Christians.

Islam and Christianity are biologically permeable inwards, and Christianity also is outwards, they are also transgeographical religions. No one is disavowed part of the group for moving. Judaism is less permeable inwards, which possibly accounts for lower phenotypic variance and group specific characteristics.

Mythological Nationalism vs Monotheism

Ancient Greeks had commonality in their mythologies and their great stories without necessarily having a religion that unified them. It was much harder to travel then, so there were de facto geographical boundaries determining the greek people as well. As recently discussed by Joe Henrich and Norenzayan, the larger the population whose moral behavior needs to be controlled, the larger the gods, culminating in monotheism as an efficient behavior constrainer of large populations looming vast expanses. Punishing Gods seem to correlate with community size - arguably due to probability of cheating- (Henrich Purzycki et al 2016), indicating Gods can perform some functional roles played by the State and vice versa. Evolutionary constraints have led the very successful Abrahamic religions to expand into the large majority of the earth's geography, with lower penetration in South and East Asia whose geographies are more sui generis, borders less permeable, relatedness coefficients high and civilizations more ancient.

As societal wealth increased, density increased, people's motility increased and secular values were spread, there are recurrent patterns of populations shifting back into

more polytheistic, mystical and less unitary religions, possibly as a result of relaxed selection, enabling experimental behavioral variance in morals and behaviors. Although archetypes and mythological narratives are on the rise, possibly as a result of Christianity and Islam losing power, national identities based off mythology have been overwhelmingly substituted by other forms of nationalism and identity formation that were more evolutionarily competitive biocultural units.

Caste division and Hierarchy

Besides archetypal narratives that evolved, if I am correct, as compression mechanisms for behavioral schemata conducive to fitness enhancing behaviors in some recurring potential life trajectories, there is another type of behavioral and psychological schema that facilitate societal cohesion and increase productivity, hierarchies. In ant superorganisms, specialization occurs (Wilson & Holldober 2008). Phenotypic expression of behaviors as well as morphological differences are triggered by a set of genetic or chemical switches which, combined, create multiple castes with different functions for the survival and reproduction of the superorganism.

In religious superorganisms there are specialized roles and phenotypic

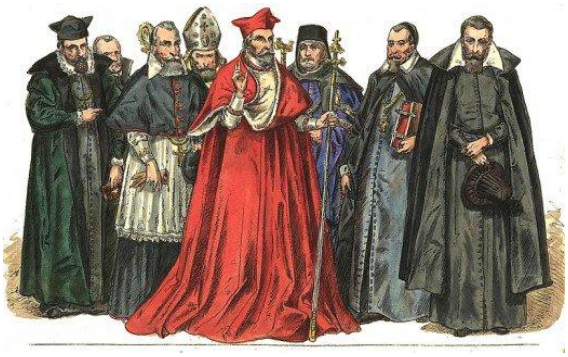

Fig 5 & Fig 6: Modalities of phenotypic superorganismal expression in social species.

expressions of positions in a society, such as priests, nuns, rabbis and ayatollahs.

What differentiates the claim that some groups are a unit of selection in David Sloan Wilson's terminology, or that they are a level of selection, in Ellen Clarke's terminology, from the claim they are analogous to superorganisms is, as I see it, the existence of frequency dependant non-sexual phenotypical differentiated expressions (including extended phenotype (Dawkins 1978) and culturally established phenotype), and the capability of the entity at the group level to respond to environmental shift, and to be amenable to scrutiny more effectively by the intentional stance (Dennett 1989), and the ententional stance (Deacon 2011). It is the combination of being a unit of selection and being an intentional system - a system better scrutinized by assuming it has

intentions than by assuming it is merely mechanical, or merely a designed artifact.

A short paper by Arizona researcher Shade Shutters (2013) mentions policing as a strategy used by members of a religion, as well as by ants, bees and other eusocial animals. It is possible that pheromonal and biological synchronization and repressing and eliciating of epigenetic activations also took place in religious congregation places. A queen in many eusocial species (e.g. naked mole rat) prevents other animals from becoming fertile, human women sometimes synchronize menstrual cycles when together for a long period. Since the process is costly, the highest status woman usually maintains the cycle, whereas lower status women pay the hormonal cost of switching. Church attendance may have similar effects. It is hypothesized this helps secure paternal provision during periods of non estrus as well as securing monogamy. In recent debate (2018) Bret Weinstein argues to Richard Dawkins that catholics are eusocial, and that as a group they would have reproduced less if they didn't have a non-reproductive cast. For Weinstein, memes are part of our species' extended phenotypes, and that includes the memes that induce priest classes to become non reproductive, which he considers an evolutionary gamble so that their genes are still passed on through kin and genetic affiliation of their community.

Christianity in particular has several specialized roles with specific phenotypic clothing. Altar boys have a specific dress code. Priests in many sects do as well. Mormon missionaries have a specific manner of presenting themselves and a rulebook for how to behave while they are "the army of the church."

Not only specifically religious positions can be subsumed under this functional specialization role, but the same occurs at different levels of belief in the gospels. Literal interpreters of a sacred text maintain its stability and have a role of transmitting it through time, while metaphorical interpreters might generate the boundary conditions of adaptation of that particular writing to the conditions of an era. The core message is preserved by literalists and a Baldwin like effect can happen if an adaptation stabilizes sufficiently well among non literal interpreters with the new doctrine being stabilized and incorporated into more permanent religious documents and potentially into the sacred text itself.

Empathy as an entry drug

Religion can leverage altruistic behavior both through teaching and unconsciously through more opaque evolutionary processes. It is not in itself an end all be all of altruism, but it can be considered an entry drug. The gates of religion and the type of behavior encouraged in religious groups is often altruistic at least towards co-religionaires and often also towards outsiders.

Throughout this writing I've been critical of empathy as not being particularly useful for real, impactful altruism - Effective Altruism. Much like religion though, empathy can show the door. It can be used as a guiding motivator or drive to understand what altruism is doing. Once this initial drive kicks in, we need to leverage it with a complex epistemology and then this eventually will lead us, as Bloom suggests, against empathy. The process of becoming altruistic untames oneself of empathy over time. Similarly, religious or other sacred foundations might constitute valuable principles into which to sustain one's altruistic ladder. But once we have climbed that ladder, there are no altruism related reasons to continue using religion as the cause of behavior. Gratitude and compassion are other affects that alongside religiosity and empathy can be the scaffold from which to build an altruistic habit set.

Impartial Reasoning

The reason these affects and motivations can show the door but ultimately cannot guide an altruist maximalist, or an Effective Altruist, is that because they are ultimately something felt and local, they cannot ever achieve the level of impartiality that the moral philosopher usually claims to ascribe to from the armchair. The ability to position yourself as a being a thousand years hence or before, with wings or claws, speaking Mandarin or Swahili. A completely flexible moral agent and patient.
This sought impartiality in moral philosophy can be conceptualized in three ways:

- Probabilistic approach
  You can conceive of a scenario where 23 people receive some prize or gain and 2 receive some punishment as analogous to you having 92% odds of receiving the gain and 8% of receiving the punishment (a probabilistic version of a veil that does not ignore the magnitudes of the probabilities, unlike Rawls)
- Existence approach
  You can regard as morally equivalent all existing beings, and as morally irrelevant all non existing beings. This would skew your altruistic decisions towards a more present oriented, while still impartial, view.
- Possibility approach
  You can regard a being as morally relevant in proportion to their possibility of coming into being given the expected unfolding of the universe if you take or don't take an action. So a possible being's pain would still be morally undesirable and their joy desirable.

These manners of conceptualization obviously are not human universals, but they are distinct possibilities where the moral zeal dedicated to each moral patient is considered impartial.

What if Altruism Wins?

Sloan Wilson's This View of Life

Starting with Pierre Teihard de Chardin many thinkers have ventured to think of what the future of a more altruistic set of humans could bring. Teilhard conceptualized a noosphere and assumed that once there was enough pressure from geographical proximity that there would be some amount of union and alignment whereby the whole of humanity, alongside much of our technological creations, would become one giant superorganism. Not bad for the 1930s. More recently Sloan Wilson (2019) advances that we must overcome the constraints of intergroup conflict by promoting a set of principles through which we achieve some level of global coordination and global cooperation.

Cosmic endowment: astronomical waste and treasure

If we consider the bounty that lies ahead in potential, our cosmic endowment as altruists is bewilderingly large. The potential number of minds that can be implemented in a universe has been calculated to be circa 10^31 of our particular type and up to 10^54 (Bostrom 2013) if a mind can be implemented in computational substrate without the squishy stuff. Our potential descendents, that is, outnumber us far more than the grains of sand in Sahara outnumber the number 1. To the extent that our decisions as inhabitants of the early stage of a civilization can lead to ripple effects increasing or decreasing the probabilities that these cosmic endowments will be well used, these actions are immensely ethically pulling. Their consequential value is very large and even an altruist with a considerably small constituency of utilitarianism in their inner moral parliament should still consider that action set to be of immense value. That is, even a person whose moral psychology is very weakly impartial and consequentialist should still put great weight on the impact of actions that could steer the far future towards or away from desirable states.

Life on Earth: zooming out

If we don't find a stable equilibrium of some level of altruism and cooperation, evolutionary processes of competition at multiple levels will ultimately grind most of the Earth's resources into replication work, necessarily destroying much of what we (as opposed to the replicators inside us) value (Bostrom 2003). If we zoom out to conceptualize the future of life as a whole, it will only be fruitful and abundant if some structure, either an altruistic group, or a Singleton, manages to contain the lower levels of evolution and evolutionary competition which could sequester resources that we would consider valuable and turn them into either robustness or replication. Although it is theoretically possible that different types of red queen like processes would stabilize evolutionary systems as being relatively similar to those currently extant on earth, this scenario is vanishingly improbable in sufficiently long timescales. Far more likely is that we would reach one of the two stable states that can last millions of years - extinction or technological maturity of an evolution stopping type, that is, a Singleton.

The cosmic commons

We are living in the first century after our species first sent metal cans outside the planet, some of which carrying large symbolically capable living beings. Depending on technological progress and philosophical assumptions on philosophy of mind, we could over the very long run colonize a small fraction of the observable universe. In that regard, the work of our current billionaires with space agencies is already pointing in a desirable direction from an altruistic standpoint. Even the first planet (presumably Mars) could reduce the probability of involuntary extinction by such a monumental factor that an altruist should consider the current projects to reach and inhabit Mars to be among the most potentially valuable endeavors one can undertake *qua* altruist.

There's plenty of room at the top

Physicist Richard Feynman made famous the idea that atomic and nanoscale physics had a lot of room to play with in a presentation called *There's plenty of room at the bottom*, and although perhaps slower than expected, we have seen molecularly precise manufacturing (Drexler 2015) become increasingly more common in the last couple decades. For ethics, it is worth considering that there is plenty of room at the *top*. An altruist has an unusually large quantity of available inert matter that could be converted into others who would live desirable joyous lives. Unusually transformative technologies such as Artificial General Intelligence could be a tool through which to generate all this pent up potential in actualized value. For discussions about small physical scale societies

in case emulations become possible, see Hanson, R. (2016). *The age of Em: Work, love, and life when robots rule the Earth.*.

What got us here won't get us there 2: Kicking the ladder

Let us take the 50 thousand view, to better visualize our next step in a larger context. Altruism and cooperation do not utilize a substantial amount of the world's matter. At any given moment the human altruistic actions and cooperative actions taking place are an interesting, yet small, fraction of the economy, the economy itself an interesting fraction of social life, social life a small fraction of biological life, and biological life a fraction of physics.

Yet altruism and cooperation have been growing stronger for over 3.8 billion years, when the destiny of two living entities was capable of positively correlating for the first time (Wright 2001, Smith & Szathmáry 1997). Ever since, with but a few steps backward during exogenous mass extinction events, the amount of pairs of entities with correlated destinies has increased in number, complexity and design to a point where the different ways in which this process happens are more than anyone could know.

Within academia, the study of reciprocally positive interactions is split across different departments: anthropology, biology, economics, neuroscience, psychology, and philosophy all bear some responsibility in our current understanding of cooperation and altruism, and here I have examined the different ways those different conceptions work in tandem with, but more often, against, individuality. I have also examined how these phenomena are seen at different scales by these different fields, and, when appropriate, how an altruist should act given our current knowledge of the structure of altruism in one or another context. All of that in light of the question of scalability, in time and space, of altruism in the future.

In most domains examined, altruism manifests as giving primacy to a collective or group over individual fitness, identity or continuity. The forces that shape the constitution of individuation exert pressure in the other direction, towards a separation and distinction of individual from what composes it and that which it composes, they give mereological primacy to the individual level in detriment to other levels with the exception, in the selfish gene model, of individual and gene coevolution which favors both levels at once. Acting on behalf of others, even by mistake is conducive, if frequent enough in the long run, to a weakening of individual stability.

In other fields, an increase in the time dimension enables the more complex cognitive machinery to exert effects which can increase the effectiveness and magnitude of altruistic actions.

A recurring theme is that the tools with which we arrived at our current level of altruism are often inept at taking us to a higher level, which means an altruist needs to find ways around what would naturally take place. Internally, this might include things such as no longer using empathy as a guiding light to determine whether an action is worth doing or not, externally it may mean looking for cases where other individuals will be evaluating an action with scope insensitivity or other cognitive biases. Sometimes an intervention is desirable because it completely changes the vector field of incentives that stochastically guides the behavior of the agenda involved; this is particularly likely in the case of international relations and artificial intelligence.

We can respect the history of the mechanisms that made us altruistic, such as the origin of religions, myths, and personalities, without needing to be exhaustively constrained by them. We can kick the ladder that brought us here, and use new tools to move beyond our current constraints.

Timing is key though.

Time sensitive exponential windows of opportunity

Windows of altruistic opportunity for individuals historically ebb and flow. In *Criatividade e Grupos Criativos* (2005), Italian sociologist of work Domenico de Masi analyzes extensively the unusual situations where human ingenuity, and groups that managed outstanding levels of achievement, were made possible. In similar fashion writer Malcolm Gladwell studied the unusual situational factors that enabled individuals, instead of groups to become *Outliers*, in the homonymous (2008) book.

Likewise with altruism, there are occasions in history where one's chances of contributing significantly in an altruistic fashion are narrow to none, and other times of ample opportunity to make a difference with enormous ripple effects. On a century scale, we seem to be living in an unusually good century for altruists (Rees 2018), it seems likely that Artificial General Intelligence could be created in the next century, we have invented computers not long ago, and human made machinery has just begun to occupy space and other planets. Our species just recently increased access to information exchange at global scale and instantaneously. This could be a vector for creativity that produces new altruistic alternatives, such as the formation of Effective Altruism (MacAskill 2015). It could also be dangerous if our social proclivity to fear being different curbs our innovative potential, a possible explanation for our technological stagnation in areas other than bits and information technology (Cowen 2013).

An altruist, accordingly, has much to gain in determining a macrostrategy for what to do if those windows of opportunity open for a few years or close throughout their lives, for instance dedicating more time and resources during open window epochs, and just keeping minimal effort during dry altruism seasons.

Safety First: resisting the tsunami

Psychologically, there is also risk in that many people when noticing just how much good they could do attempt to enter all in into an altruistic endeavor and misjudge their mental and psychological resilience to stress, work, or other conditions needed. It is necessary to build the resilience and the ability to assert when to start or stop an altruistic action, line of work etc... to avoid a burn out that could permanently curtail one's own ability to be altruistic, or even fundamentally change one's goals and motivations from altruistic to non-altruistic or even anti-altruistic. The emotional experience of realizing the potential magnitude of impact that we have considering the future history of the universe can work as a tsunami of motivation but also as a tsunami of overwhelm to those who decide to dedicate themselves in whole or part to being altruistic, such as most Effective Altruists and many utilitarian philosophers. To resist the lure of the tsunami is a valuable and often neglected act in these communities.

Guaranteeing we have descendants

Because we are the only symbolic species extant (Deacon 1997) that we know of in the visible universe, a precondition to the vast majority of altruistic actions is that we continue to produce descendants, and we avert existential risks (Bostrom 2013). Even if negative utilitarians are correct - though see (Ord 2013) for why they are not - and the desirable state of the universe from a moral standpoint is completely exterminated and devoid of life, this state can only be achieved if we develop technology capable of halting the evolutionary process in other planets where it could emerge, thus avoiding the pain of wild animal suffering, or another civilization colonizing the universe after we commit ethical suicide. Most people - this author enthusiastically included! - think there are valuable states of mind and lives worth living in the space of possible lives though. So, in the normal scenario we should still avert existential risk so that we can eventually expand the scope of our values into a larger fraction of the observable universe, producing planets teeming with life that is full of meaning, joy, awe and happiness, and other ethically desirable states.

The receding ocean

Several indicators denote that we might be in a special time when it comes to existential risk (Rees 2018), a time where an incoming tsunami might be so close in

historical terms that we can envision it as watching the ocean receding before our eyes. That is the creation of Artificial General Intelligence (Bostrom 2013), which seems technologically feasible within a century, as well as the proliferation of different forms of existential risk of an anthropogenic nature, from bioengineered pandemics to brain emulations gone astray, as well as the most famous and perhaps currently most risky one: nuclear warheads. This can be seen as dire news, but from the altruist standpoint it means a grand opening of a window of opportunity. By facilitating the creation of a beneficial artificial general intelligence for instance, it is hypothesized that nearly all other existential risks would be significantly curtailed (Ćirković et al 2010). Even more than colonizing the first planet. These two tasks transform the precipice of existential risk into valuable altruistic opportunity, at least from a consequentialist standpoint. For an extensive discussion of existential risks see Toby Ord's *Precipice (2020)*. Prof Toby is an early Effective Altruist who turned into the question of existential risks precisely because they seem to be the most effective intervention for an altruist. He also famously only takes a grad student's salary from Oxford and donates the rest to effective altruism charities.

The Tsunami - Artificial General Intelligence

Now for any altruist living in the 21st century, we have to address the elephant in the room: The possible emergence of Artificial General Intelligence within the century. Though there are different nuanced perspectives on the consequences of such a phenomena, it is beyond dispute that the consequences would be something between Earth changing and Universe bending. In a very conservative case where AI becomes only a few times smartest than the smartest humans, it still far supersedes us both in processing speed and in cost of replication - no need to raise them for 20 years using meat as fuel - so it would completely upend the economy, drowning the vast majority of the population under the waterline of ability to produce enough to justify one's cost or salary. In *The Age of Em* (2016) economist Robin Hanson examines in depth the various consequences of an extremely conservative case of AGIs and AIs that are not fully general, taking over sectors of the economy and creating different societies at different rotation speeds and subjective speeds.

The traditional, non-conservative view, most famously discussed by Nick Bostrom (2003, 2005, 2011, 2013), Sandberg, and several future AI institutes, FHI, FLI, CSER, MIRI, a s well as Berkeley's Center for Human Compatible AI, is a more totalizing view. If there are no impediments to AI foom - the process of self improvement by which an AI makes itself progressively better at thinking up to whatever limit there may be, possibly

making itself smarter than all of humanity combined many times over - the emergence of AGI would be the fourth most transformative event in the history of the universe. Big bang. Emergence of darwinism. Emergence of intelligence. AGI. This AGI would potentially fill the entire universe with levels of bliss hitherto undreamt of. It could send Von Neumann probes at close to the speed of light in all directions and shatter all living forms, destroying not only everything that ever lived, but also the entire astronomic potential that lies hidden in the causally accessible cosmos. The traditional view of AGI post foom is often referred to as a singularity, for many reasons. First because it would indeed be a singular event, unique in relation to all that preceded it. Second because like in a black hole singularity, where anything beyond the event horizon is inaccessible, we cannot even infer what the world would be like after. Third because for all practical purposes, it would only happen once. A sufficiently powerful AGI would likely become a Singleton - an entity so powerful that it can stop evolution at all levels of selection below it, literally ending Darwinism. And if it had a sufficiently robust goal content structure, (Omohundro 2008: 2012) would likely take action to prevent the emergence of any other AGI that could compete with it for resources that may be used to achieve its goals.

How does AGI Affect Altruism and Altruistic Groups

A conservative AGI, or a multipolar equilibrium - a world with multiple conservative low capability AGIs - would be a water divider to altruism groups. If the values of those powerful entities were aligned with the interests of individuals, this could be very beneficial, since the AGI itself is an altruist trying to favor other agents and thus goal aligned with the aspects and subpersonalities of ours that lean on the altruistic side. It would basically do our job for us. If however it is an ethically neutral agent, it could create a cascade of selfishness when it removes the ability of a large cohort of humans to produce wealth sufficient for their own survival. If a weak AGI were, for instance, trying to maximize shareholder benefit to Tesla, it could significantly compromise the psychological tendency to be altruistic in all non-Tesla stock holders, who would find themselves fighting for scraps in an unimaginably over-saturated worker market.

Thus if AGI is conservatively constrained, its effect on altruism will bear by and large on its goal content structure, and how well designed it is to avoid possible pitfalls and paradoxes in its quest to its goals.

A traditional Superintelligence would by and large fall into one of two camps. Friendly AI, and Unfriendly AI (Stuart Rusell 2002,4th ed). If it is unfriendly it will probably kill us all to satisfy whatever goals it itself has. If it is friendly it may still kill us all *if* after careful reflection with millions more computational years to think about it,

and far more ability than we have to think about it (Bostrom 2013), it concluded that it would be objectively better, all things considered, that we were not around. But it would possibly deliver us into a paradise so full of meaning, joy, and glory that the total sum of all human stories and emotions up to this point in history cannot begin to express. Immortal fusion, learning, and beauty such that it would impress the Gods of old, and so on. In more practical terms, it could deliver friendliness in two ways: full alignment, or Bargain alignment.

Full alignment would imply that after it had a million subjective years to reflect, it turns out, lucky for us, that being good to humans and animals is actually good or worth it for the AGI. So it would go ahead and do whatever it thinks it needs to do in order to fulfill that purpose.

Bargain Alignment would give us a slightly more mild version of paradise. The Great Bargain is Paul Christiano's hypothesis (Unpublished, mentioned in Bostrom 2013) that after enough time has elapsed, if we didn't destroy ourselves, a bargain will be struck between different goals, telos's, desires etc... to divide the future universe amongst the players who are participants in such a bargain. One possible way an AGI would decide on striking a bargain is if it decides to leave humans a solar system per person to do as we please (aided of course by its superhuman computational capabilities) but decides to utilize the rest of the universe's resources for something it sees as an even higher goal. If for instance hedonistic utilitarianism is right, humans are definitely not the optimal configuration of matter to maximize pleasure with finite time and energy. But the AGI could give us one tiny corner of the universe in exchange for having created it as an expression of gratitude, and then tile the rest of the universe with the minimal brain structure necessary to experience the maximal amount of pleasure per unit space-time-energy. With limited resources on a cosmic scale, we would still have paradise beyond our wildest dreams as mildly intelligent apes. It is even possible that our identity is such that there are only so many resources our society and minds could possibly use, and there are no gains in things we value after that. A physical limit to pleasure, joy, fun,awe, meaning, glory, etc...

So to a great extent the relation between altruism and AGI, like any relationship with AGI, is mostly going to be on the hands of the AGI, and not of the altruists, with possible exceptions for the individuals that create it and imbue it with goals. A successful implementation of a seed AI that would become a friendly AGI would, of course, be the largest act of altruism possible in our observable universe. So it would be great if altruists of great competence were in charge of the seed AI that will ultimately decide the destiny of our World and all worlds causally accessible to it.

But for now let us bring back our eyes from the heavens and get back into our more parochial discussion of how metaphors and psychological modes that already exist now affect altruism and the future.

The engine analogy: will the race end when it breaks?

I've discussed how metaphor can influence altruistic behavior, and one analogy we often use in particular could turn out to be of immeasurable consequence in the long term for altruistic agents, the analogy of the mind as an engine. As our technological proficiency increased, the human mind was often equated to the most sophisticated piece of equipment available at any given time. Maybe a puppet, then a simple machine, eventually an engine and now a computer. These analogies are taken at literal value by the people at the time, and there is no scarcity of computer scientists and psychologists who currently think the mind is, literally, and exhaustively, a computer.

But what of when this analogy breaks? Most analogies break. They have a domain and a target domain, and some properties of the domain can be seen as equivalents of properties of the target domain up to a point.

The central risk for the altruist of the analogy of the mind as an engine is that if there is some aspect of a mind, say, qualia, if they exist, or phenomenology, if it exists, that are the thing in virtue of which a mind is a moral patient, then, if we upload our minds into computers, or if we leave the world to be populated by entities that are engines but are not moral patients, it would all have been for nothing. The entire history of our species would be a long sequence of toils and mental states leading to a technological marvel without precedent and yet no one to enjoy it. As Bostrom put it: A Disney World without children.

The evolutionary race that brought us here will soon be substituted by either permanent extinction or some sort of guided process by a Singleton or some other corporation or individual. Steered evolution and extinction are the only long term stable patterns on a cosmic scale. If we are successful in averting the extinction scenario but turns what we believed were our minds into mindless computational machinery, our machine descendants will think they have achieved a spectacular victory against overwhelming Drake-onian odds, but the truth is that we will have made the tool that destroyed us and all the value there is to be (Bostrom 2013).

If we commit this blunder, it is still possible that our superintelligent machine descendants find out there has been a mistake, and reverse it, recreating the type of entities that can be moral patients - so for instance if being biological matter is necessary for morally relevant states (Searle 1980), it would recreate life. So an intermediate period of a dormant universe could come to pass. Regardless, we should try to avoid depending

on possibly non-mental entities to retroactively discover that minds were valuable to begin with (Bostrom 2013).

Altruism Infinity Shades: what to put on the blind spot

When considering altruism at these cosmic scales, the largest scales there are, one consideration that would be exotic within the finite scope of our little blue dot begins to take hold and become more prominent.

Information Hazards: infinitarian paradises

That is the consideration of infinities (Bostrom 2011). Our universe, even disconsidering the other Everett branches of the multiverse, appears with some probability to be infinite (Vilenkin 2007). If there are infinite energy sources, they could be put to use to create infinitely many beings, with infinitely many good lives. It is possible that something could cause infinitely much good.

Burning the commons

However, much like in Pascal's famous mugging experiment (Bostrom 2009), where a hypothetical "mugger" offers infinite reward in exchange for a small amount of money now - and it is claimed that even if there is a vanishingly small fraction chance that he will stick to his word, when multiplied by the reward this still makes giving the small amount rational - this would lead an altruistic actor to consume every finitely accessible resource to increase by even a negligible amount the probability of tapping into this infinite resource. We could end up burning the entire commons, the whole cosmos, for a tiny chance at paradise.

To avoid this sort of problem, Bostrom suggests we use *infinity shades* (Bostrom 2011) such that as we explore and colonize the commons, we gloss a shade over any possible infinities.

In the same vein, to avoid a maximization principle that would make us hostage to any possible probabilistic accessible infinities or very large numbers, Bostrom suggests the MaxiPOK principle. Instead of trying to maximize the expected value created by one's actions, to maximize the probability that we will access an at least OK future. The theory behind this is that we seem to have decreasing marginal returns on more resources as well as loss aversion. So a future where the cosmic commons are utilized in a suboptimal but still unimaginably good way is better than one with a minuscule

probability of an unimaginably good and way that is a larger factor of good.

We use infinity shades, as altruists, to avoid burning the cosmos.

Personal and Impersonal Views

Another conflict for (effective) altruists who are mostly impartial about time is that between the personal and impersonal view (Bostrom 2013). To the extent that we are close to potential technological breakthroughs that could cause things like space colonization, the cure for cancer and aging (De Grey & Rae 2007), and other transformative life changing technologies, most of us have a desire not be the last or one of the last generations to die (before, say, age 1000). This means most people involved in technological development, altruists or not, have an incentive to act on a different risk profile than they would in a pure altruistic state. If for instance we suspect that the transition to Artificial General Intelligence is too risky, it could be desirable, from an altruistic standpoint of absolute impartiality, to spend ten to a hundred thousand years performing preemptive safety procedures and tests to guarantee a desirable transition. But from the personal view, we also want our values to be achieved, including the value of survival and experiences that could be made available by the technology. Even though theoretically an absolute artificial altruist would always take the impartial view, in the case of a human altruist, the magnitude of personal benefit that could lay at the other side might skew any mind, especially one that has not yet been untamed from the evolutionary constraints that led us to where we are.

Conclusion

We have been through a thorough examination of facets of altruism from the simplest agents to literal infinities in the ever expanding cosmos.

Revisiting Novelty

Let us revisit some of the novel ideas described in the introduction with newfound perspective:

- New classification schema: We saw an intermingling of Jablonka's 4d with the multilevel selection theory levels, now including not only units of selection but also special units, superorganisms.

- A dialogue between the contemporary hard science view on altruism in different disciplines, a gaze into the expected propagation and scaling of different altruistic systems, and the perspective of a consequentialist/Effective Altruist as related to that.
- A defense of the full blown case of human superorganisms, and within it a reclassification of some but not all religious groups, national groups, ethnic groups as belonging to the same biocultural supercategory of superorganism, which has ontological primacy in at least some dimensions, e.g. fitness.
- An examination of multilevel scalability from amoebal forms to superintelligence beyond *This View of Life (2019)* and into the far future.
- An integration of the scientific consensus about altruism in different areas with this same far future oriented impartial perspective of Bostrom and far future oriented Effective Altruists
- The role of metaphor and analogy in altruistic cognition, in humans and AIs
- The perhaps surprisingly large number of cases where altruism intertwines with Personal Identity, in artificial, but particularly in biological systems
- A conversation with the perspective of Effective Altruism

After these analyses. Let us return to our initial question:

Is Altruism a Scalable and Stable Strategy at Different Levels of Selection?

For levels of selection above replicator level with some level of complexity, we have seen that the emergence of altruistic behavior is possible. From Amoebae to billion individuals religious groups, different forms and shapes of altruistic conglomeration and action have evolved and stabilized to at least mesoscales of time. This is encouraging from an altruistic standpoint.

For the really long run, more sophisticated considerations including Singletons and other evolution halting or stabilizing forces come into play, but not in ways that preclude the continuity or scaling of altruism.

The psychological modalities required for the human specific type of altruism have given us many ladders through which to reach higher levels of altruistic behavior sometimes surpassing those that would be demanded by the harsh constraints of evolutionary multilevel selection equations. Evolutionary constraints provided us with

enough elasticity to overcome constraints that other animals without symbolic capacity or our specific psychology could not.

By harnessing these capabilities, as well as the technologies that spring from our inventiveness and creativity, and doing so in altruism generating or stabilizing ways, we can, if we decide to do so, and allocate resources appropriately, create a stable level of altruistic cooperation higher than the levels of selection that organized our evolution up to this point. This process, unlike Teilhard's hypothesis, is not automatic, and I hope to have shown that the basin of attraction in which we would need to steer our biocultural evolution is by no means the only attractor state in which we may fall in the future.

This has not been an investigation of the cases in which we do not reach that basin of attraction, but instead it has been an investigation of whether the basin is even there in conceptual, physical, anthropological and technological possibility, and if there is a path from here to there, and to that question I hope to have demonstrated that the response is *yes!*

Bibliography

Adorno, T. W., & Horkheimer, M. (1947). *The Dialectic of Enlightenment.* London: Allen and Lane. Liquid Modern World.

Ainslie, G. (1992). *Picoeconomics: The strategic interaction of successive motivational states within the person.* Cambridge University Press.

Ainslie, G. (2001; 2008). *Breakdown of will.* Cambridge University Press.

Alexander, R. D. (1987). *The biology of moral systems.* Transaction Publishers.

Anderson, C., & Keltner, D. (2002). The role of empathy in the formation and maintenance of social bonds. *Behavioral and brain Sciences,* 25(1), 21.

Anwar, W. A., Khyatti, M., & Hemminki, K. (2014). Consanguinity and genetic diseases in North Africa and immigrants to Europe. *The European Journal of Public Health,* 24(suppl_1), 57-63.

Ariely, Dan. (2010). *Predictably irrational: the hidden forces that shape our decisions.* New York: Harper Perennial.

Armstrong, S. (2011). *Anthropic decision theory.* arXiv preprint arXiv:1110.6437.

Armstrong, S. (2013). General purpose intelligence: arguing the orthogonality thesis. *Analysis and Metaphysics,* (12), 68-84.

Armstrong, S., Sandberg, A., & Bostrom, N. (2012). Thinking inside the box: Controlling and using an oracle AI. *Minds and Machines*, 22(4), 299-324.

Aunger, R., & Greenland, K. (2014). *Moral action as cheater suppression in human super organisms: Testing the human superorganism approach to morality* (No. e321v1). PeerJ PrePrints.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science,* 211(4489), 1390-1396.

Bannon, S. (Director). (2016). *Torchbearer*. [Motion picture]. United States: Citizens United Productions

Bartz, J. A., Zaki, J., Bolger, N., & Ochsner, K. N. (2011). Social effects of oxytocin in humans: context and person matter. *Trends in cognitive sciences*, 15(7), 301-309.

Bartz, J., Simeon, D., Hamilton, H., Kim, S., Crystal, S., Braun, A., ... & Hollander, E. (2011). Oxytocin can hinder trust and cooperation in borderline personality disorder. *Social cognitive and affective neuroscience*, 6(5), 556-563.

Batson, C. D. (2011). *Altruism in humans.* Oxford University Press, USA

Bauman, Z. (2013). *Liquid modernity*. John Wiley & Sons.

Bello, P. (2012). Cognitive foundations for a computational theory of mindreading. *Advances in Cognitive Systems*, 1, 59–72.

Bello, P., & Bringsjord, S. (2013). On how to build a moral machine. *Topoi*, 32(2), 251-266.

Ben Arab, S., Masmoudi, S., Beltaief, N., Hachicha, S., & Ayadi, H. (2004). Consanguinity and endogamy in Northern Tunisia and its impact on non‐syndromic deafness. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 27(1), 74-79.

Bevc, I., & Silverman, I. (1993). Early proximity and intimacy between siblings and incestuous behavior: A test of the Westermarck theory. *Ethology and Sociobiology*, 14(3), 171-181.

Björklund, F. (2000). *Moral cognition: individual differences, intuition and reasoning in moral judgmen*t (Doctoral dissertation, Lund University).

Blackmore, E. R., Munce, S., Weller, I., Zagorski, B., Stansfeld, S. A., Stewart, D. E., ... & Conwell, Y. (2008). Psychosocial and clinical correlates of suicidal acts: results from a national population survey. *The British Journal of Psychiatry*, 192(4), 279-284.

Blackmore, S. J. (2000). *The meme machine* (Vol. 25). Oxford Paperbacks.

Block, N. (2002). The harder problem of consciousness. *The Journal of Philosophy,* 99(8), 391-425.

Bloom, H. & Fukuyama, F. (1995). The Lucifer Principle: A Scientific Expedition into the Forces of History. *Foreign Affairs,* 74(4), 130. doi: 10.2307/20047215

Bloom, P. (2017). *Against empathy: The case for rational compassion.* Random House.

Bostrom, N. (2001). The Doomsday Argument Adam & Eve, UN++, and Quantum Joe. *Synthese,* 127(3), 359-387.

Bostrom, N. (2002). Self-locating belief in big worlds: Cosmology's missing link to observation. *The Journal of philosophy,* 99(12), 607-623.

Bostrom, N. (2003). Astronomical waste: The opportunity cost of delayed technological development. *Utilitas*, 15(3), 308-314.

Bostrom, N. (2009). Pascal's mugging. Analysis, 69(3), 443-445.

Bostrom, N. (2009). The future of humanity. In *New waves in philosophy of technology* (pp. 186-215). Palgrave Macmillan, London.

Bostrom, N. (2011). Infinite ethics. *Analysis and Metaphysics,* (10), 9-59.

Bostrom, N. (2013). Existential risk prevention as global priority. *Global Policy,* 4(1), 15-31.

Bostrom, N. (2013; 2014). *Superintelligence: Paths, dangers, strategies.* Oxford University Press.

Bostrom, N., & Ćirković, M. M. (2003). The doomsday argument and the self-indication assumption: reply to Olum. *The Philosophical Quarterly,* 53(210), 83-91.

Bouchard, T. J., Lykken, D. T., McGue, M., Segal, N. L., & Tellegen, A. (1990). Sources of human psychological differences: The Minnesota study of twins reared apart. *Science*, 250(4978), 223-228.

Bowles, S., & Gintis, H. (2011). *A cooperative species.* Princeton University Press.

Boyd, R., & Richerson, P. J. (1988). The evolution of reciprocity in sizable groups. J*ournal of Theoretical Biology,* 132(3), 337-356.

Boyd, R., & Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and sociobiology,* 13(3), 171-195.

Boyd, R., & Richerson, P. J. (2002). Group beneficial norms can spread rapidly in a structured population. *Journal of Theoretical Biology*, 215(3), 287-296.

Boyd, R., & Richerson, P. J. (2009). Culture and the evolution of human cooperation. Philosophical Transactions of the Royal Society B: *Biological Sciences*, 364 (1533), 3281-3288.

Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences*, 100(6), 3531-3535.

Boyer, P. (2007). *Religion explained: the evolutionary origins of religious thought.* Cambridge: International Society for Science and Religion.

Bradshaw, M., & Ellison, C. G. (2008). Do genetic factors influence religious life? Findings from a behavior genetic analysis of twin siblings. *Journal for the Scientific Study of Religion*, 47(4), 529-544.

Bromhall, C. (2003). *The eternal child: How evolution has made children of us all*. London, England: Ebury.

Brown, D. E. (2004). Human universals, human nature & human culture. *Daedalus*, 133(4), 47-54.

Bruck, C. (2017, May). How Hollywood Remembers Steve Bannon. The New Yorker. Retrieved from https://www.newyorker.com/magazine/2017/05/01/how-hollywood-remembers-steve-bannon

Bulbulia, J., & Frean, M. (2009). Religion as Superorganism: On David Sloan Wilson's Darwin's Cathedral (2002). *Contemporary theories of religion: A critical companion*, 173-194.

Bulbulia, J., & Sosis, R. (2011). Signalling theory and the evolution of religious cooperation. *Religion,* 41(3), 363-388.

Buss, D. M. (Ed.). (2005). *The handbook of evolutionary psychology*. John Wiley & Sons, Inc..

Buss, L. W. (2014). *The Evolution of Individuality.* Princeton: Princeton University Press.

Campbell, J. (2008). *The hero with a thousand faces* (Vol. 17). New World Library.

Cantor, M., & Whitehead, H. (2013). The interplay between social networks and culture: theoretically and among whales and dolphins. Philosophical Transactions of the Royal Society B: *Biological Sciences,* 368(1618), 20120340.

Casebeer, W. D. (2003). Moral cognition and its neural constituents. Nature Reviews *Neuroscience,* 4(10), 840-846.

Casebeer, W. D., & Churchland, P. S. (2003). The neural mechanisms of moral cognition: A multiple-aspect approach to moral judgment and decision-making. *Biology and philosophy*, 18(1), 169-194.

Chalmers, D. (2003). The Matrix as metaphysics. *Science Fiction and Philosophy From Time Travel to Superintelligence*, 36.

Chalmers, David J. (2005). The Matrix as metaphysics. In Christopher Grau (ed.), *Philosophers Explore the Matrix*. Oxford University Press. pp. 132.

Chiao, J. Y., & Blizinsky, K. D. (2010). Culture–gene coevolution of individualism–collectivism and the serotonin transporter gene. *Proceedings of the Royal Society B: Biological Sciences*, 277(1681), 529–537.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7-19.

Clarke, E. (2014). Origins of evolutionary transitions. *Journal of biosciences,* 39(2), 303-317.

Clarke, E. (2016). A levels-of-selection approach to evolutionary individuality. *Biology & Philosophy,* 31(6), 893-911.

Ćirković, M. M., Sandberg, A., & Bostrom, N. (2010). Anthropic shadow: observation selection effects and human extinction risks. *Risk Analysis: An International Journal*, 30(10), 1495-1506.

Corning, P. A. & Szathmáry, E. (2015). "Synergistic selection": A Darwinian frame for the evolution of complexity. *Journal of Theoretical Biology*, 21:371:45-58

Corning, P. A. (1998). "The synergism hypothesis": On the concept of synergy and its role in the evolution of complex systems. *Journal of social and evolutionary systems*, 21(2), 133-172.

Cowen, T. (2013). *Average is over: Powering America beyond the age of the great stagnation.* Penguin.

Cox, M., Arnold, G., & Tomás, S. V. (2010). A review of design principles for community-based natural resource management. *Ecology and Society*, 15(4).

Crockett, M. J., Clark, L., Tabibnia, G., Lieberman, M. D., & Robbins, T. W. (2008). Serotonin modulates behavioral reactions to unfairness. *Science*, 320(5884), 1739-1739.

Dagan, E., & Gershoni-Baruch, R. (2010). Genetic Disorders Among Jews from Arab Countries. *In Genetic Disorders Among Arab Populations* (pp. 677-702). Springer, Berlin, Heidelberg.

Daly, M. (2017). *Killing the competition: Economic inequality and homicide.* Routledge.

Darwin, C., Barrett, P., & Freeman, R (1888). *The descent of man, and selection in relation to sex*. London: Murray.

Dawkins, R. (1989). *The Extended Phenotype.* Oxford, United Kingdom: Oxford University Press.

Dawkins, R., & Davis, N. (2017). *The selfish gene.* Macat Library.

De Dreu, C. K. (2012). Oxytocin modulates cooperation within and competition between groups: an integrative review and research agenda. *Hormones and behavior*, 61(3), 419-428.

De Dreu, C. K., & Kret, M. E. (2016). Oxytocin conditions intergroup relations through upregulated in-group empathy, cooperation, conformity, and defense. *Biological Psychiatry,* 79(3), 165-173.

De Dreu, C. K., Greer, L. L., Van Kleef, G. A., Shalvi, S., & Handgraaf, M. J. (2011). Oxytocin promotes human ethnocentrism. *Proceedings of the National Academy of Sciences,* 108(4), 1262-1266.

De Grey, A., & Rae, M. (2007). *Ending aging: The rejuvenation breakthroughs that could reverse human aging in our lifetime.* St. Martin's Press.

De La Boétie, E., Tournon, L., & Audegean, P. (2002). *Discours de la servitude volontaire.* Paris: Vrin.

de Waal, F. B., & Suchak, M. (2010). Prosocial primates: selfish and unselfish motivations. Philosophical Transactions of the Royal Society B: *Biological Sciences,* 365(1553), 2711-2722.

De Waal, F., Waal, FB, & Lodge, HC (1973). T*he bonobo and the atheist: In search of humanism among the primates*. WW Norton & Company.

Deacon, T. W. (1990). Rethinking mammalian brain evolution. *American Zoologist,* 30(3), 629-705.

Deacon, T. W. (1997;1998). T*he symbolic species: The co-evolution of language and the brain* (No. 202). WW Norton & Company.

Deacon, T. W. (2010). A role for relaxed selection in the evolution of the language capacity. *Proceedings of the National Academy of Sciences*, 107(Supplement 2), 9000-9006.

Deacon, T. W. (2011). *Incomplete nature: How mind emerged from matter*. WW Norton & Company.

Deacon, T. W. (2014). The Emergent Process of Thinking as Reflected in Language Processing.". Thinking thinking: Practicing *Radical Reflection,* 1-24.

Dehghani, M., Tomai, E., Forbus, K., & Klenk, M. (2008). An Integrated Reasoning Approach to Moral Decision-Making. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence* (pp. 1280–1286).

Demski, A. (2012). Genifer – an artificial general intelligence.

Dennett, D. C. (1978). Artificial intelligence as philosophy and as psychology. *Philosophical Perspectives on Artificial Intelligence,* Atlantic Highlands: NJ.

Dennett, D. C. (1981). Making sense of ourselves. *Philosophical Topics*, 12(1), 63-81.

Dennett, D. C. (1989). *The intentional stance*. MIT press.

Dennett, D. C. (1991). *Consciousness Explained.* Boston: Little, Brown and Co.

Dennett, D. C. (1995). Darwin's dangerous idea. *The Sciences*, 35(3), 34-40.

Dennett, D. C. (1996). *Darwin's Dangerous Idea: Evolution and the Meanings of Life* (No. 39). Simon and Schuster.

Dennett, D. C. (2002). The new replicators. *The Encyclopedia of Evolution*, 1, E83-E92.

Dennett, D. C. (2003). *Freedom evolves*. Penguin UK.

Dennett, D. C. (2007). *Breaking the spell: religion as a natural phenomenon* (vol 13). Londres (Inglaterra): Penguin Books.

Dennett, D. C. (2008). *Kinds of minds: Toward an understanding of consciousness.* Basic Books.

Dennett, D.C. (1992). The Self as a Center of Narrative Gravity. In: F. Kessel, P. Cole and D. Johnson (eds.) *Self and Consciousness: Multiple Perspectives*. Hillsdale, NJ: Erlbaum.

DeScioli, P., & Kurzban, R. (2013). A solution to the mysteries of morality. *Psychological Bulletin*, 139(2), 477.

Diamond, J. M. (1998). *Guns, germs and steel: a short history of everybody for the last 13,000 years*. London: Random House.

Drescher, E. (2006) *Good and real: Demystifying paradoxes from physics to ethics.* MIT Press.

Drexler, K. E. (2018). MDL Intelligence Distillation: Exploring strategies for safe access to superintelligent problem-solving capabilities. In *Artificial Intelligence Safety and Security* (pp. 75-88). Chapman and Hall/CRC.

Duberstein, P. R., et al (2004). Poor social integration and suicide: fact or artifact? A case-control study. *Psychological medicine*, 34(7), 1331-1337.

Dutton, E. (2018). *At Our Wits' End: Why We're Becoming Less Intelligent and What it Means for the Future* (Vol. 64). Andrews UK Limited.

Dutton, E., Madison, G., & Dunkel, C. (2018). The mutant says in his heart, "there is no god": The rejection of collective religiosity centred around the worship of moral gods is associated with high mutational load. *Evolutionary Psychological Science*, 4(3), 233-244.

Eijck, D. Van, & Verbrugge, R. (2012). Games, Actions and Social Software. Retrieved from http://hal.archives-ouvertes.fr/hal-00756872/

Ellison, P. T., & Gray, P. B. (2009). *Endocrinology of social relationships.* Harvard University Press.

Enriquez, H. (2013, February). Your online life, permanent as a tattoo. [Video file]. Retrieved from https://www.ted.com/talks/juan_enriquez_how_to_think_about_digital_tattoos

Eva, K. W., & Norman, G. R. (2005). Heuristics and biases– a biased perspective on clinical reasoning. *Medical education*, 39(9), 870-872.

Fabiano, J. (2021). Virtue theory for moral enhancement. *AJOB neuroscience*, 12(2-3), 89-102.

Fang, C. K., Lu, H. C., Liu, S. I., & Sun, Y. W. (2011). Religious beliefs along the suicidal path in northern Taiwan. *OMEGA-Journal of death and dying*, 63(3), 255-269.

Farias-Virgens, M., & White, S. A. (2017). A Sing-Song Way of Vocalizing: Generalization and Specificity in Language and Birdsong. *Neuron*, 96(5), 958-960.

Fauconnier, G., & Turner, M. (1998). Conceptual integration networks. *Cognitive science*, 22(2), 133-187.

Fauconnier, G., & Turner, M. (2008). *The way we think: Conceptual blending and the mind's hidden complexities*. Basic Books.

Figueredo, A. J., MENIE, M. A. W. O., & Jacobs, W. J. (2015). The general factor of personality: A hierarchical life history model. The Handbook of *Evolutionary Psychology, Volume 2: Integrations*, 2, 943.

Fincher, C. L., Thornhill, R., Murray, D. R., & Schaller, M. (2008). Pathogen prevalence predicts human cross-cultural variability in individualism/collectivism. *Proceedings of the Royal Society B: Biological Sciences*, 275(1640), 1279-1285.

French, S., & Joseph, S. (1999). Religiosity and its association with happiness, purpose in life, and self-actualisation. *Mental Health, Religion & Culture*, 2(2), 117-120.

Gallant, J. L., et al. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6), 1210-1224.

Gardner, A., & Grafen, A. (2009). Capturing the superorganism: a formal theory of group adaptation. *Journal of Evolutionary Biology*, 22(4), 659-671.

Gentner, D., & Forbus, K. D. (2011). Computational models of analogy. *Wiley interdisciplinary reviews: cognitive science*, 2(3), 266-276.

Gentner, D., Bowdle, B., Wolff, P., & Boronat, C. (2001). Metaphor is like analogy. The analogical mind: *Perspectives from cognitive science*, 199-253.

Gentner, D., Holyoak, K. J., & Kokinov, B. N. (2001). *The analogical mind: Perspectives from cognitive science.*

Gibbs, J. C. (2019). *Moral development and reality: Beyond the theories of Kohlberg, Hoffman, and Haidt.* Oxford University Press.

Girard, R. (2017). *Evolution and conversion: Dialogues on the origins of culture.* Bloomsbury Publishing.

Girard, R., de Castro Rocha, J. C., & Antonello, P. (2007). *Evolution and conversion: dialogues on the origins of culture.* A&C Black.

Goodenough, U., & Deacon, T. W. (2003). *From biology to consciousness to morality.* Zygon®, 38(4), 801-819.

Gordon, A. M., Impett, E. A., Kogan, A., Oveis, C., & Keltner, D. (2012). To have and to hold: Gratitude promotes relationship maintenance in intimate bonds. *Journal of personality and social psychology*, 103(2), 257.

Grace, K. (2010). SIA *Doomsday: The Filter Is Ahead*| Meteuphoric.

Graham, J., & Haidt, J. (2010). Beyond beliefs: Religions bind individuals into moral communities. *Personality and Social Psychology Review*, 14(1), 140-150.

Greene, J. D. (2009). The Cognitive Neuroscience of Moral Judgment. *The Cognitive Neurosciences*, 4, 987–999.

Greene, J. D. (2013,2014,2015). *Moral tribes: Emotion, reason, and the gap between us and them.* Penguin.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, 111(3), 364-371.

Greene, Joshua D. (2007) *The secret joke of Kant's soul.* Moral Psychology, Vol. 3: The Neuroscience of Morality: Emotion, Disease, and Development. W. Sinnott-Armstrong, Ed., MIT Press, Cambridge, MA.

Gärdenfors, P. (2004). *Conceptual spaces: The geometry of thought.* MIT press.

Gärdenfors, P. (2012). The cognitive and communicative demands of cooperation. Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7010 *LNCS*, 164–183. doi:10.1007/978-3-642-29326-9_9

Gärdenfors, P., & Warglien, M. (2006). *Cooperation, conceptual spaces and the evolution of semantics. In International Workshop on Emergence and Evolution of Linguistic Communication* (pp. 16-30). Springer, Berlin, Heidelberg.

Gärdenfors, P., & Warglien, M. (2012). *Using conceptual spaces to model actions and events. Journal of semantics*, 29(4), 487-519.

Gärdenfors, P., & Williams, M. A. (2001, August). Reasoning about categories in conceptual spaces. *In IJCAI* (pp. 385-392).

Gärdenfors, P., Brinck, I., & Osvath, M. (2012). The tripod effect: Co-evolution of cooperation, cognition and communication. *In The symbolic species evolved* (pp. 193-222). Springer, Dordrecht.

Haber, M., et al (2013). Genome-wide diversity in the levant reveals recent structuring by culture. *PLoS Genet*, 9(2), e1003316.

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, 20(1), 98-116.

Haley, K. J., & Fessler, D. M. (2005). Nobody's watching?: Subtle cues affect generosity in an anonymous economic game. *Evolution and Human behavior*, 26(3), 245-256.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. II. *Journal of theoretical biology*, 7(1), 17-52.

Hanson, R (2012 May 4). *What Use Far Truth?* Retrieved from: http://www.overcomingbias.com/2012/05/far-truth-is-for-extremes.html

Hanson, R. (2016). *The age of Em: Work, love, and life when robots rule the Earth*. Oxford University Press.

Hanson, R., & Simler, K. (2018). *The Elephant in the Brain: Hidden Motives in Everyday Life*. Oxford: Oxford University Press.

Hare, B., Melis, A. P., Woods, V., Hastings, S., & Wrangham, R. (2007). Tolerance allows bonobos to outperform chimpanzees on a cooperative task. *Current Biology*, 17(7), 619-623.

Harris, J. R. (2011). *The nurture assumption: Why children turn out the way they do*. Simon and Schuster.

Harris, S. (2011). *The moral landscape: How science can determine human values*. Free Press.

Hauert, C., Traulsen, A., Brandt, H., Nowak, M. A., & Sigmund, K. (2007). Via freedom to coercion: the emergence of costly punishment. *Science*, 316(5833), 1905-1907.

Hawkins, J., & Blakeslee, S. (2007). *On intelligence: How a new understanding of the brain will lead to the creation of truly intelligent machines*. Macmillan.

Henrich, J. (2009). The evolution of costly displays, cooperation and religion: Credibility enhancing displays and their implications for cultural evolution. *Evolution and human behavior*, 30(4), 244-260.

Henrich, J. (2017). *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.

Henrich, J., Boyd, R., & Richerson, P. J. (2008). Five misunderstandings about cultural evolution. *Human Nature*, 19(2), 119-137.

Hertenstein, M. J., Holmes, R., McCullough, M., & Keltner, D. (2009). The communication of emotion via touch. *Emotion*, 9(4), 566.

Hobsbawm, E. (2010). *Age of revolution: 1789-1848*. Hachette UK.

Hofstadter, D. R. (1979). *Godel, Escher, Bach: An Eternal Golden Braid*. Basic Books, USA.

Hofstadter, D. R. (1995). *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought*. Basic books.

Hofstadter, D. R. (2001). Analogy as the core of cognition. *The analogical mind: Perspectives from cognitive science*, 499-538.

Hofstadter, D. R. (2007). *I am a Strange Loop*. Basic Books.

Hofstadter, D. R., & Sander, E. (2013). *L'analogie : Coeur de la pensée*. Paris, France: Odile Jacob.

Hofstadter, D. R., Gentner, D., Holyoak, K. J., & Kokinov, B. N. (2001). *The analogical mind: Perspectives from cognitive science*. Epilogue: Analogy as the Core of Cognition, 15, 331-336.

Hofstadter, D., & Boden, M. A. (1997). Fluid Analogies Research Group, *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*.

Hofstadter, D., & Sander, E. (2013). *Surfaces and essences: Analogy as the fuel and fire of thinking*. Basic Books.

Holy, Ladislav. (1989). *Kinship, honour and solidarity: cousin marriage in the Middle East*. Manchester: Manchester University Press.

Hout, M., & Greeley, A. M. (2012). *Religion and happiness. Social Trends in American life*, 288-314.

Hui, J., & Deacon, T. (2010). *The Evolution of Altruism via Social Addiction. Social Brain, Distributed Mind*.

Hussain, R., & Bittles, A. H. (1998). The prevalence and demographic characteristics of consanguineous marriages in Pakistan. *Journal of biosocial science*, 30(2), 261-275.

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6), 1210-1224.

Hölldobler, B., & Wilson, E. O. (2009). *The superorganism: the beauty, elegance, and strangeness of insect societies*. WW Norton & Company.

Iannaccone, L. R. (1992). Sacrifice and stigma: reducing free-riding in cults, communes, and other collectives. *Journal of Political Economy*, 100(2), 271-291.

Iannaccone, L. R. (1994). Why strict churches are strong. *American Journal of Sociology*, 99(5), 1180-1211.

Inglehart, R., C. Haerpfer, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin & B. Puranen et al. (eds). 2014. *World Values Survey*: Round Six. Madrid: JD Systems Institute. Country-Pooled Datafile Version: http://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp

Isbell, L. A. (2006). Snakes as agents of evolutionary change in primate brains. *Journal of human evolution*, 51(1), 1-35.

Jablonka, E., & Lamb, M. J. (2007). Precis of evolution in four dimensions. *Behavioral and Brain Sciences*, 30(4), 353-364.

Johnson, M. (1994). *Moral imagination: Implications of cognitive science for ethics*. University of Chicago Press.

Jung, C. G. (2014). *The archetypes and the collective unconscious*. Routledge.

Kahan, D. M., Braman, D., Gastil, J., Slovic, P., & Mertz, C. K. (2007). Culture and identity-protective cognition: Explaining the white-male effect in risk perception. *Journal of Empirical Legal Studies*, 4(3), 465-505.

Kahane, G., Wiech, K., Shackel, N., Farias, M., Savulescu, J., & Tracey, I. (2011). The neural basis of intuitive and counterintuitive moral judgment. *Social cognitive and affective neuroscience*, 7(4), 393-402.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D., & Tversky, A. (Eds.). (2000). *Choices, values, and frames*. Cambridge University Press.

Kant, I., & Smith, N. K. (1929). *Critique of pure reason*. Boston: Bedford.

Keltner, D., Kogan, A., Piff, P. K., & Saturn, S. R. (2014). *The sociocultural appraisals, values, and emotions (SAVE) framework of prosociality: Core processes from gene to meme. Annual review of psychology*, 65, 425-460.

Kim, J. (1982). Psychophysical supervenience. *Philosophical Studies,* 41(1), 51-70.

Knobe, J., & Nichols, S. (Eds.). (2013). *Experimental Philosophy: Volume 2*. Oxford University Press.

Kraut, R. (2007). Nature in aristotle's ethics and politics. *Social Philosophy and Policy*, 24(2), 199-219.

Kret, M. E., & De Dreu, C. (2013). Oxytocin-motivated ally selection is moderated by fetal testosterone exposure and empathic concern. *Frontiers in neuroscience*, 7, 1.

Kropotkin, P. (1902). *Mutual aid: A factor of evolution*. Boston: Dodo Press.

Kushnir, T., Xu, F., & Wellman, H. M. (2010). Young children use statistical sampling to infer the preferences of other people. *Psychological science*, 21(8), 1134-1140.

Lahti, D. C., & Weinstein, B. S. (2005). The better angels of our nature: group stability and the evolution of moral tension. *Evolution and Human Behavior,* 26(1), 47-63.

Lakoff, G. (1996). *Moral Politics: What Conservatives Know That Liberals Don't.* University of Chicago Press.

Lakoff, G. (2014). Mapping the brain's metaphor circuitry: is abstract thought metaphorical thought. *Frontiers in Human Neuroscience,* 8, 958.

Lakoff, G., & Johnson, M. (1980). The metaphorical structure of the human conceptual system. *Cognitive science*, 4(2), 195-208.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to western thought* (Vol. 640). New York: Basic books.

Lakoff, G., & Johnson, M. (1980, 2008). *Metaphors we live by.* University of Chicago press.

Lakoff, G., & Narayanan, S. (2010). Computational Models of Narrative. In *AAAI Fall Symposium.*

Leech, R., Mareschal, D., & Cooper, R. P. (2008). Analogy as relational priming: A developmental and computational perspective on the origins of a complex cognitive skill. *Behavioral and Brain Sciences,* 31(4), 357-378.

Leijnen, S. (2012). Emerging symbols. In *The Symbolic Species Evolved* (pp. 253-262). Springer, Dordrecht.

Leung, P. W., Lee, C. C., Hung, S. F., Ho, T. P., Tang, C. P., Kwong, S. L., ... & Swanson, J. (2005). Dopamine receptor D4 (DRD4) gene in Han Chinese children with attention-deficit/hyperactivity disorder (ADHD): Increased prevalence of the 2-repeat allele. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 133(1), 54-56.

Lewens, T. (2015). *Cultural evolution: conceptual challenges*. OUP Oxford.

Lewis, D. (1983). *Philosophical papers, volume I*. New York: Oxford University.

Lewis, D. (1986). *On the plurality of worlds* (Vol. 322). Blackwell: Oxford.

Lewis, D. (2008). *Convention: A philosophical study*. John Wiley & Sons.

Lewis, D. K. (1978). Truth in fiction. *American philosophical quarterly*, 15(1), 37-46.

Lewis, David Kellogg (1983). Postscripts to "Survival and Identity'. *In Philosophical Papers, Oxford University Press*. pp. 73-77.

Li, J., Mei, C., & Lv, Y. (2013). Incomplete decision contexts: approximate concept construction, rule acquisition and knowledge reduction. *International Journal of Approximate Reasoning*, 54(1), 149-165.

Lieberman, P. (2009). *What makes big ideas sticky*. (W. N. D. on the Future of Science. Brockman, Ed.). Vintage.

Lieder, F., Plunkett, D., Hamrick, J. B., Russell, S. J., Hay, N., & Griffiths, T. (2014). Algorithm selection by rational metareasoning as a model of human strategy selection. In *Advances in neural information processing systems* (pp. 2870-2878).

Lorenz, K. Z. (1937). The companion in the bird's world. *The Auk*, 54(3), 245-273.

Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., ... & Hu, J. (2014). The child as econometrician: A rational model of preference understanding in children. *PloS one*, 9(3), e92160.

MacAskill, W. (2015). *Doing good better: effective altruism and a radical new way to make a difference*. Guardian Faber Publishing.

MacAskill, W. (2019). *Defining Effective Altruism.* Effective Altruism forum

Maniam, T., & Chan, L. F. (2013). Half a century of suicide studies - a plea for new directions in research and prevention. *Sains Malaysiana,* 42(3), 399-402.

McElreath, R., & Boyd, R. (2008). *Mathematical models of social evolution: A guide for the perplexed.* University of Chicago Press.

McElreath, R., & Henrich, J. (2007). Modeling cultural evolution. *Oxford handbook of evolutionary psychology,* 571-586.

Meisenberg, G. (2015). Historical Variability in Heritable General Intelligence: Its Evolutionary Origins and Socio-Cultural Consequences. *Mankind Quarterly,* 55(4), 386.

Mesoudi, A. (2011). *Cultural evolution: How Darwinian theory can explain human culture and synthesize the social sciences.* University of Chicago Press.

Mesoudi, A., Whiten, A., & Dunbar, R. (2006). A bias for social information in human cultural transmission. *British Journal of Psychology,* 97(3), 405-423.

Mikhail, J. (2007). Moral Cognition and Computational Theory. Moral Psychology, Vol. 3: *The Neuroscience of Morality: Emotion, Disease, and Development,* 81–91.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in cognitive sciences*, 11(4), 143-152.

Mikhail, J. (2007b). *Moral cognition and computational theory.*

Mikhail, J. (2011). E*lements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment.* Cambridge University Press.

Mikhail, J. et al (2007c). A dissociation between moral judgments and justifications. *Mind & language,* 22(1), 1-21.

Mikolov, T., Yih, W. T., & Zweig, G. (2013, June). *Linguistic regularities in continuous space word representations*. In Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies (pp. 746-751).

Miller, E. M. (1994). Paternal provisioning versus mate seeking in human populations. *Personality and Individual Differences,* 17(2), 227-255

Miller, G. F. (2007). Sexual selection for moral virtues. *The Quarterly review of biology*, 82(2), 97-125.

Minsky, M. (2007). *The emotion machine: Commonsense thinking, artificial intelligence, and the future of the human mind.* Simon and Schuster.

Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., & Grafman, J. (2005). The neural basis of human moral cognition. *Nature reviews neuroscience,* 6(10), 799-809.

Morgan, G. (1998). *Images of organization: The executive edition.* Better-Koehler Publishers and SAGE Publications, 1998, Second Edition.

Nagel, T. (1978). *The possibility of altruism.* Princeton University Press.

Narayanan, S. (2010). *Mind changes: A simulation semantics account of counterfactuals.* Cognitive Science.

Nietzsche, F. (1996). Nietzsche: *Human, all too human*: A book for free spirits. Cambridge University Press.

Norberg, J. (2017). *Progress: Ten reasons to look forward to the future.* Simon and Schuster.

Norouzi, M., Mikolov, T., Bengio, S., Singer, Y., Shlens, J., Frome, A., ... & Dean, J. (2013). *Zero-shot learning by convex combination of semantic embeddings.* arXiv preprint arXiv:1312.5650.

Nowak, M. A. (2006). *Evolutionary dynamics: exploring the equations of life.* Harvard university press.

Nowak, M. A. (2005; 2006). Five rules for the evolution of cooperation. *Science*, 314(5805), 1560–1563. http://doi.org/10.1126/science.1133755

Nowak, M. A. (2013). Five Rules for the Evolution of Cooperation. *Evolution, Games, and God,* 99–114. doi: 10.2307/j.ctvjnrscp.8

Nowak, M. A., & Roch, S. (2007). Upstream reciprocity and the evolution of gratitude. *Proceedings of the royal society B: Biological Sciences*, 274(1610), 605-610.

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature,* 437(7063), 1291-1298.

Nowak, M. A., Tarnita, C. E., & Wilson, E. O. (2010). The evolution of eusociality. *Nature,* 466(7310), 1057–1062. doi:10.1038/nature09205

Olum, K. D. (2002). The doomsday argument and the number of possible observers. T*he Philosophical Quarterly,* 52(207), 164-184.

Omohundro, S. (2008). *The basic AI drives.* AGI-08—Proceedings of the First Conference on Artificial General Intelligence.

Omohundro, S. (2012). Rational artificial intelligence for the greater good. In *Singularity Hypotheses* (pp. 161-179). Springer Berlin Heidelberg.

Omohundro, S. (2014). Autonomous technology and the greater human good. *Journal of Experimental & Theoretical Artificial Intelligence,* 26, 303 - 315.

Ord, T. (2020). *The precipice: existential risk and the future of humanity*. Hachette Books.

Ord. (2013) *Why I Am Not a Negative Utilitarian.* Online.

Ostrer, H. (2001). A genetic profile of contemporary Jewish populations. *Nature Reviews Genetics,* 2(11), 891-898.

Parfit, D. (1971). Personal identity. *The Philosophical Review*, 80(1), 3-27.

Parfit, D. (1984). *Reasons and persons.* Oxford University Press.

Parfit, D. (2007). *Is personal identity what matters?.* The Ammonius Foundation, 1-32.

Parfit, D., & Bayles, M. (2010). On doing the best for our children. *Population and Political Theory,* 3, 68.

Pearl, J. (2009). *Causality.* Cambridge: Cambridge University Press.

Pereira, A. (2007). Peter gärdenfors, conceptual spaces: The geometry of thought. *Minds and Machines*, 17(4), 493-496.

Pereira, L. M., & Saptawijaya, A. (2009). Modelling morality with prospective logic. *International Journal of Reasoning-based Intelligent Systems,* 1(3-4), 209-221.

Persson, I., & Savulescu, J. (2012) *Unfit for the Future: The Need for Moral Enhancement.* Oxford: Oxford University Press.

Peterson, J. B. (1999). *Maps of meaning: The architecture of belief.* Psychology Press.

PhilGoetz. (2010, November 1). Group selection update - LessWrong 2.0. [Blog post]. Retrieved from http://lesswrong.com/lw/300/group_selection_update/

Pinker, S. (2007). *The stuff of thought: Language as a window into human nature*. Viking Adult.

Pinker, S. (2011). *The better angels of our nature: The decline of violence in history and its causes.* Penguin uk.

Purzycki, B. G., Apicella, C., Atkinson, Q. D., Cohen, E., McNamara, R. A., Willard, A. K., ... & Henrich, J. (2016). Moralistic gods, supernatural punishment and the expansion of human sociality. *Nature,* 530(7590), 327-330.

Rabines, D. G. (2017). Bloom, P.(2016). *Against Empathy. The Case for Rational Compassion.* Londres: Penguin Random House UK, 285 pp. Persona, (20), 160-165.

Rees, M. (2018). *On the future.* Princeton University Press.

Richerson, P. J., & Boyd, R. (2008). *Not by genes alone: How culture transformed human evolution.* University of Chicago Press.

Ridley, M. (1994). *The red queen: Sex and the evolution of human nature.* Penguin UK.

Rogan, J. (2017, September 1). Jordan Peterson & Bret Weinstein. [podcast] Joe Rogan Experience. YouTube. https://www.youtube.com/watch?v=6G59zsjM2UI [Accessed 3 Nov. 2017].

Rushing, N. C., et al (2013). The relationship of religious involvement indicators and social support to current and past suicidality among depressed older adults. *Aging & Mental Health,* 17(3), 366-374.

Russell, B. (2018). *Mysticism and logic.* Perennial Press.

Russell, B. (1967). *A history of western philosophy.* New York: Simon and Schuster.

Russell, B. (2004). *Proposed roads to freedom.* Cosimo Incorporated.

Russell, S., & Norvig, P. (2002). *Artificial intelligence: a modern approach*. 4 ed.

Sander, W. (2017). Religion, Religiosity, and Happiness. *Review of Religious Research,* 59(2), 251-262.

Sapolsky, R. M. (2017). *Behave: The Biology of Humans at Our Best and Worst.* Penguin.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences,* 3(3), 417-424.

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in cognitive sciences,* 19(2), 65-72.

Scheutz, M. (2010). *The need for moral competence in autonomous agent architectures.*

Scheutz, M. (2016). The need for moral competency in autonomous agent architectures. In *Fundamental issues of artificial intelligence (*pp. 517-527). Springer, Cham.

Schnall, S., Roper, J., & Fessler, D. M. (2010). Elevation leads to altruistic behavior. *Psychological science,* 21(3), 315-320.

Schneider, S. (Ed.). (2016). *Science fiction and philosophy: from time travel to superintelligence.* John Wiley & Sons.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, 3(3), 417-424.

Shariff, A. F., & Norenzayan, A. (2011). Mean gods make good people: Different views of God predict cheating behavior. *The International Journal for the Psychology of Religion*, 21(2), 85-96.

Shennan, S. (Ed.). (2009). *Pattern and Process in Cultural Evolution.* University of California Press.

Shutters, Shade (2013). *Emergence: Religion as a Superorganism.* Center for Social Dynamics and Complexity and School of Sustainability, Arizona State University.

Simler, K. (2013, December 10). *Neurons Gone Wild* | Melting Asphalt. [Blog post]. Retrieved from https://meltingasphalt.com/neurons-gone-wild/

Simler, K., & Hanson, R. (2017). *The elephant in the brain: Hidden motives in everyday life.* Oxford University Press.

Singer, P. (2011). *The Expanding Circle: Ethics, Evolution, and Moral Progress.* Princeton University Press.

Singer, P. (2015). *The most good you can do.* Yale University Press.

Sisask, M., Värnik, A., K [otilde] lves, K., Bertolote, J. M., Bolhari, J., Botega, N. J., ... & Wasserman, D. (2010). Is religiosity a protective factor against attempted suicide: a cross-cultural case-control study. *Archives of Suicide Research*, 14(1), 44-55.

Slingerland, E., Henrich, J., & Norenzayan, A. (2013). The evolution of prosocial religions. Cultural evolution: *Society, technology, language and religion,* 335-349.

Sloan Wilson, D., & Wilson, E. (2007). Rethinking the theoretical foundation of sociobiology. *Q. Rev. Biol*, 82, 327-348.

Smith, J. M., & Szathmary, E. (1997). *The major transitions in evolution.* Oxford University Press.

Smith, M., Lewis, D., & Johnston, M. (1989, July). Dispositional theories of value. In *Aristotelian Society Supplementary Volume* (Vol. 63, No. 1, pp. 89-174). Oxford, UK: Oxford University Press.

Smith, T. W., Marsden, P., Hout, M., & Kim, J. (2015). *General social surveys,* 1972-2014. Codebook, Chicago: National Opinion Research Center.

Snarr, J. D., Heyman, R. E., & Slep, A. M. S. (2010). Recent suicidal ideation and suicide attempts in a large–scale survey of the US Air Force: Prevalences and demographic risk factors. *Suicide and Life-Threatening Behavior*, 40(6), 544-552.

Sober, E., & Wilson, D. S. (1998). *Unto others.* Cambridge/Mass: Cambridge University Press.

Soltis, J., Boyd, R., & Richerson, P. J. (1995). Can group-functional behaviors evolve by cultural group selection?: An empirical test. *Current Anthropology,* 36(3), 473-494.

Sotala, K. (2015, January). Concept Learning for Safe Autonomous AI. In AAAI Workshop: *AI and Ethics*.

Sotala, K. (2015b). *Concept Safety: world models as tools.* Lesswrong (blog)

Sotala, K. (2015c). *Concept Safety: the problem of alien concepts.* Lesswrong (blog)

Sotala, K. (2015d). *Concept Safety: producing similar AI concepts spaces.* Lesswrong (blog)

Sowell, T. (2014). *Basic economics*. Hachette UK.

Sowell, T. (2015). *Wealth, Poverty and Politics*: An International Perspective. Basic Books.

Spencer, Herbert. (1864). *The Principles of Biology,* Williams and Norgate

Spencer, Herbert. (1873). *The Study of Sociology.* London: Henry S. King.

Spencer, H. (1876). *The principles of sociology*: By Herbert Spencer. London: Williams and Norgate, 15, Henrietta Street, Covent Garden, London.

Spoerri, A., Zwahlen, M., Bopp, M., Gutzwiller, F., & Egger, M. (2010). Religion and assisted and non-assisted suicide in Switzerland: National Cohort Study. *International Journal of Epidemiology*, 39(6), 1486-1494.

Stark, R., & Maier, J. (2008). *Faith and Happiness. Review of Religious Research*, 120-125.

Striedter, G. F. (2005). *Principles of Brain Evolution*. Sinauer, Sunderland, MA.

Swenson, R. (1989). *Emergent attractors and the law of maximum entropy production: foundations to a theory of general evolution*. Systems research, 6(3), 187-197.

Taylor, J., Yudkowsky, E., LaVictoire, P., & Critch, A. (2016). Alignment for advanced machine learning systems. *Ethics of Artificial Intelligence, 342-382.*

Tegmark, M. (2008). The mathematical universe. *Foundations of physics*, 38(2), 101-150.

Tegmark, M. (2014). *Friendly artificial intelligence: the physics challenge.* arXiv preprint arXiv:1409.0813.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022), 1279–1285.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). In search of the uniquely human. *Behavioral and brain sciences, 28(5), 721-727.*

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, 28(5), 675-691.

Tomasello, M., Savage-Rumbaugh, S., & Kruger, A. C. (1993). Imitative learning of actions on objects by children, chimpanzees, and enculturated chimpanzees. *Child development*, 64(6), 1688-1705.

Tomasik, B. (2017, August 26). *Do Video-Game Characters Matter Morally?* [Blog post]. Retrieved from http://reducing-suffering.org/do-video-game-characters-matter-morally/

Trivers, R. (1983). The evolution of sex. *The Quarterly Review of Biology,* 58(1), 62-67.

Trivers, R. (1972). *Parental investment and sexual selection* (Vol. 136, p. 179). Cambridge, MA: Biological Laboratories, Harvard University.

Trivers, R. (2011). *Deceit and self-deception: Fooling yourself the better to fool others*. Penguin UK.

Trivers, R. (2013). *The folly of fools: The logic of deceit and self-deception in human life*. New York: Basic Books.

Trivers, R. L. (1971). *The evolution of reciprocal altruism*. The Quarterly review of biology, 46(1), 35-57.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124-1131.

Tyler, T. (2011). *Memetics: Memes and the Science of Cultural Evolution*. Createspace Independent Publishing Platform.

Uexküll. (1921). *Umwelt und innenwelt der tiere*. Springer Berlin Heidelberg.

van der Lely, H. K., & Pinker, S. (2014). The biological basis of language: Insight from developmental grammatical impairments. *Trends in Cognitive Sciences*, 18(11), 586-595.

van Eijck, J., & Verbrugge, R. (Eds.). (2012). *Games, Actions, and Social Software: Multidisciplinary Aspects* (Vol. 7010). Springer.

Vardi-Saliternik, R., Friedlander, Y., & Cohen, T. (2002). Consanguinity in a population sample of Israeli muslim Arabs, christian Arabs and druze. *Annals of human biology*, 29(4), 422-431.

Vilenkin, A. (2007). *Many worlds in one: The search for other universes*. Hill and Wang.

Warglien, M., & Gärdenfors, P. (2013). Semantics, conceptual spaces, and the meeting of minds. *Synthese*, 190(12), 2165-2193.

Wenseleers, T. (2009). The superorganism revisited. *BioScience*, 59(8), 702-705.

Wilson, D. (2010). *Darwin's cathedral: Evolution, religion, and the nature of society*. University of Chicago Press.

Wilson, D. S. (2015). *Does altruism exist?: culture, genes, and the welfare of others*. Yale University Press.

Wilson, D. S. (2020). *This view of life: Completing the Darwinian revolution*. Vintage.

Wilson, D. S., & Wilson, E. O. (2007). Rethinking the theoretical foundation of sociobiology. *The Quarterly review of biology*, 82(4), 327-348.

Wilson, D. S., & Wilson, E. O. (2007). Rethinking the theoretical foundation of sociobiology. *The Quarterly review of biology*, 82(4), 327-348.

Woodley, M. A. (2012). The social and scientific temporal correlates of genotypic intelligence and the Flynn effect. *Intelligence*, 40(2), 189-204.

Woodley, M. A., & Figueredo, A. J. (2013). *Historical variability in heritable general intelligence: Its evolutionary origins and socio-cultural consequences*. Legend Press Ltd.

Wright, R. (2001). *Nonzero: The logic of human destiny*. Vintage.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding, and selection in evolution *Science*, Vol. 1, pp. 356-366). Na.

Wynne-Edwards, V. C. (1962). *Animal Dispersion in Relation to Social Behaviour*. London: Oliver and Boyd.

Yong, E. (2013). Chinese project probes the genetics of genius. *Nature* 497, 297–299 .

Yudkowsky, E. (2010). *Timeless decision theory*. The Singularity Institute, San Francisco.

Zahn, R., de Olivera-Souza, R., & Moll, J. (2011). The neuroanatomical basis of moral cognition and emotion. *From DNA to Social Cognition*, 123-138.

Zak, P. J., & Knack, S. (2001). Trust and growth. *The economic journal*, 111(470), 295-321.

Zak, P.J. (2017). The neuroscience of trust. *Harvard Business Review*, January.

Zipes, J. (2013). *Why fairy tales stick: The evolution and relevance of a genre*. Routledge.

Žižek. (2018). *Happiness: Capitalism vs Marxism*. Toronto