Why Are We Conscious? A Social Scientific Explanation*

Chris Bidner† Patrick Francois‡

June 6, 2025

Abstract

Despite much scholarly attention, we still have no satisfactory account of the function of consciousness, let alone an account of the specific conditions under which it will emerge via natural selection. Our theory provides such accounts by turning away from the study of the neuroanatomy of individuals and toward the study of equilibrium incentives governing the interaction of individuals. We formalize the notion of consciousness and show how it can emerge under specific conditions, despite having no effect on production, because of its *social* role in the process of securing desirable partners.

1 Introduction

Consciousness, in the sense used here, is the existence of an *inner life*: the thoughts, sensations, and feelings that constitute 'what it is like' to be a system acting in the world.[1] For many, the motivation to study consciousness is encapsulated in David Chalmers' observation that "There is nothing that

*We are grateful for comments received at various conferences and seminars, including Chicago Becker-Friedman IOG, ThReD (MIT Sloan), New York University, University of British Columbia, and University of California San Diego.

†Department of Economics, Simon Fraser University. email: cbidner@sfu.ca

‡Vancouver School of Economics, UBC. email: Patrick.Francois@ubc.ca

[1] Although "consciousness" has many meanings in everyday usage, the definition we use is the standard concept used across academic disciplines and is based on Nagel (1974). Philosophers sometimes refer to it as phenomenal consciousness (Block, 1995), while the terms 'subjectivity' or a 'first-person perspective' are other common synonyms. The contents of consciousness are variously referred to as phenomenal experience, subjective experience, inner experience, and qualia. We formalize the concept in Section 2.

we know more intimately than conscious experience, but there is nothing that is harder to explain" (Chalmers, 1995). Among the various puzzles presented by consciousness is the question: *Why do we have an inner life at all?* Answering this central question is interesting in its own right, but also has the potential to make inroads into related puzzles.[2]

Our goal in this paper is to provide an answer to this central question by identifying a fitness value of consciousness and to derive conditions under which it will evolve. No compelling answer currently exists, despite a recent, rapidly-growing, and multi-disciplinary revival of interest in studying consciousness (Nagel, 2012). The well-known challenge here is that, once initial intuitions are interrogated and dismissed, consciousness does not seem to do anything (Blackmore & Troscianko, 2018; Dennett, 1991). After all, it is easy to imagine various cognitive capabilities that would help an agent navigate a given situation, but what *additional* benefit could possibly arise from the agent merely knowing 'what it is like' in that situation? In other words, *conditional* on the various capacities to act in a given situation, how could it possibly help to also have a subjective experience of that situation? In the words of Fodor (2004): "As far as anybody knows, anything that our conscious minds do they could do just as well if they weren't conscious."

Our approach identifies a fitness value of consciousness, one that respects existing arguments for its causal impotency, by looking in a new place; not the neuroanatomy of individuals, but the nature of equilibrium in social interactions and their implied evolutionary dynamics.

Our first step is to formalize how consciousness can be embedded in an otherwise standard economic model. We do this in section 2. Our approach is squarely aligned with the "what it is like" definition, allows for gradations of consciousness, ties subjective experience with behaviour, and is consistent with the above observations that consciousness seems to do nothing. In short, we encode the neurological activity underlying some behaviour in a cognitive state and map these states to one of $q \in \{0, 1, ...\}$ subjective states. The subjective state associated with a cognitive state is interpreted as the inner experience resulting from existing in that cogni-

[2]Among many others, these puzzles include understanding the following questions. What other systems (such as non-human animals, plants, and computers) have inner lives? Will increasingly capable mechanical systems, such as those incorporating increasingly human-like AI, develop consciousness? How do conscious experiences arise from the physical world? The first and second of these questions arise from The Problem of Other Minds and the third from The Mind-Body Problem (Churchland, 2013), and, specifically, The Hard Problem of Consciousness (Chalmers, 1995).

tive state. We allow the value of q to vary across agents and we refer to it as an agent's *consciousness type*.

Using this formulation, we proceed to analyze how consciousness could have evolved in Section 3. In line with existing arguments for why consciousness does not do anything, our formulation rules out the possibility that subjective experience has any effect on the behaviour that accompanies it. Consciousness provides an agent with information about "what it is like" to be them in various situations, but, again, this information is useless if we consider individuals in isolation. We uncover a potential role for consciousness by taking a *social* perspective.[3]

The model features agents with a two-dimensional type–economic type and consciousness type–and is social in the sense that an agent's output depends on their economic type *and* the economic type of their partner. Again, consciousness type plays no role in production. Rather, its value lies solely in attracting desirable partners. In short, differences in economic type will manifest in differences in inner lives. Observing the inner life of others would therefore allow desirable types to pair, enhancing their fitness. But *inner* lives are not observable. Instead, agents match on the basis of reports of inner lives which we call *personas*. Undesirable types can mimic desirable types, but incur a cost when doing so because, unlike desirable types, they need to learn what the inner life of a desirable type consists of. Importantly, higher consciousness types have richer inner lives and are therefore more costly to mimic.

At each point in time, agents take the two-dimensional type distribution as given and decide whether to mimic. The Bayes-Nash equilibrium of this signaling-with-matching game then determines the equilibrium payoff from adopting the persona of a desirable agent. This, in turn, determines the agents' fitness levels–output net of any mimic costs and biological costs–at that date. The relative fitness levels then drive the joint evolution of economic type and consciousness type. In the long run the distribution of consciousness type is degenerate on a single value and we are primarily interested in this value.

We find that consciousness can arise in the long run, under quite specific conditions, because it facilitates signaling. Higher consciousness types

[3]The general idea that the origins of consciousness are social in nature is not prominent in consciousness studies, but has been recognized as a possibility – e.g. Barlow (1997), who notes similar themes in the work of Nietzsche, and more recently by Humphrey (2023)) – but such work is entirely informal. Our work benefits from the standard formal tools of Economics, and, in showing that consciousness arises only under very specific conditions, we show how these tools are indispensable in reaching our conclusions.

are more costly to mimic, which means that they are more likely to pair with a desirable type in equilibrium. This will produce a fitness gain if this additional economic payoff compensates for the additional biological cost of higher consciousness. Whereas existing approaches to understanding consciousness largely involve interrogating the information-processing potential of neuronal hardware within individuals, our approach emphasizes the properties of equilibria in social interactions and how these interact with evolutionary dynamics.

The formal model yields a number of new insights into the emergence of consciousness. A necessary feature of the economic environment is that output increases in the economic type of partners but decreases in one's own economic type: i.e. a Prisoners' Dilemma (e.g. tragedies of the commons and public goods problems). Moreover, we show that only certain types of Prisoners' Dilemmas can possibly lead to non-trivial consciousness. We show that mimicking costs can't be too small nor too large, and that any parameterization of mimic costs implies an upper bound on consciousness complexity even in the absence of biological costs. Yet, we show how the model can generate arbitrarily high consciousness levels by choosing an appropriate combination of mimic and biological costs. In deriving the co-evolution of preferences and consciousness, the model draws a novel connection between consciousness and 'prosocial' preferences.

We know of no existing research that formalizes (phenomenal) consciousness, proposes a fitness value, and derives conditions under which it will evolve. Within Economics, there is no existing work on consciousness at all, yet our model draws from and extends research on signaling-with-matching and preference evolution.

Models of signaling-with-matching involve agents taking costly actions in order to enhance their matching prospects (e.g. Cole et al. (1995), Hoppe et al. (2009), Hopkins (2012), and Bidner (2010, 2014)). In our model, the costly action is reporting on the inner life of a desirable economic type, whereby the cost of the action is the amount of mimicry required and is thus higher when mimicking higher consciousness types. The addition of second dimension of type that only affects signaling costs is new, but the key difference to these models is the addition of evolutionary dynamics. That is, our model allows us to endogenize the distribution of economic types. Less obviously, an evolutionary approach eliminates the perennial issue of equilibrium selection in signaling models; positive masses of each possible type means there is no freedom to support behaviour by imposing suitable "off-equilibrium" beliefs. Moreover, existing models in this literature assume some form of complementarity between economic types (to

ensure single-crossing) but our interesting results rely on this not being the case.[4]

A standard model of preference evolution considers a one-dimensional type[5] and concludes that non-selfish preferences can emerge to the extent of assortative matching on types (see Robson and Samuelson (2011) or Alger (2023) for overviews). The assortativeness may arise from kin relationships (Kay et al., 2020), from the extent to which types are observable (Frank (1987), Dekel et al. (2007)), or may be simply taken as exogenous (Alger & Weibull, 2013). In contrast, we consider the evolution of a two-dimensional type where the new dimension (consciousness) affects incentives to mimic and therefore plays a central role in *endogenously* determining the extent of positive assortative matching. That is, being able to recount the details of a desirable agent's inner life acts as a "secret handshake" of sorts. Unlike existing work (Robson (1990), Wiseman and Yilankaya (2001)), mimics do not cause agents to abandon a secret handshake for another, leading to cycles, but, rather, they provide the evolutionary pressure for what amounts to a more elaborate handshake (i.e. additional conscious complexity).

The paper proceeds as follows. Section 2 describes how we formalize consciousness and Section 3 develops the evolutionary model of consciousness. The main results are presented in Section 4, where we present conditions under which consciousness emerges via natural selection. Section 5 discusses our findings and their relevance for the literature on consciousness. Section 6 concludes.

2 Defining Consciousness

Before analyzing the evolution of consciousness we must clearly formalize what we mean by it. Our goal is to remain faithful to the standard (but entirely informal) definition, based on Nagel (1974), that emphasizes

[4]We get single-crossing from the fact that mimic costs are not incurred by those who faithfully report their inner life. See Kartik (2009) for a (non-evolutionary) model of signaling with lying costs. Because our mimic costs are determined by consciousness type, our model can be seen as endogenizing lying costs in these sorts of models.

[5]Heller and Mohlin (2019) consider the evolution of a two-dimensional type, where the new dimension is a capacity for deception. Although not an evolutionary model, Hopkins (2014) shows how an ability to "mentalize" (form a theory of mind) enhances the fitness of altruists in a static setting. In that context, mentalizing means getting a more precise signal about the prior actions of partners in a long term relationship, and is therefore unrelated to consciousness.

the subjective and experiential nature of a conscious being. That is, we want our formalization to capture the idea that there is "something that it is like" to be a conscious being in particular states. Broadly speaking, our approach involves first endowing agents with cognitive states and then attaching subjective states to each.

Consider a standard economic model in which an agent of economic type $\theta \in \Theta$ faces a decision problem embedded in environment $\omega \in \Omega$. As usual, the economic type encapsulates the agent's preferences over outcomes, and the environment encapsulates the map from behaviour to outcomes. The agent's behaviour is then taken to be one that induces the most-preferred outcome. So far so good.

This general approach to understanding human behaviour has proved useful, in part, because it abstracts from the murky neurological details that constitute the agent's decision making process.[6] Yet, we can always identify the details of a decision making process with a *cognitive state*. In particular, we can say that cognitive state $\phi \in \Phi$ produces a behavioural response $R(\phi)$ so that the response of an agent with economic type θ to the decision problem posed in environment ω, denoted $r(\theta, \omega)$, was produced by some cognitive state in $\Phi(\theta, \omega) \equiv R^{-1}(r(\theta, \omega))$.

To be sure, this explicit consideration of cognitive states is redundant in the standard model. We introduce it only to formalize how we think about consciousness. In particular, it allows us to offer an objective, third person, description of behaviour. To introduce a subjective, first person, description of "what it is like" *for the agent* when faced with the decision problem posed in environment ω, we introduce the notions of *consciousness type* and *subjective states*. We endow agents with a consciousness type, $q \in \{0, 1, ...\}$. When an agent of consciousness type q is in cognitive state ϕ, their subjective state is given by a function $s_q : \Phi \to \{0, 1, ..., q\}$. That is, $s_q(\phi)$ is interpreted as "what it is like", from the agent's point of view, to be in cognitive state ϕ.[7] Higher values of q are therefore interpreted as the range of inner experiences available to the agent. For instance, an agent with $q = 0$ is non-conscious because every cognitive state maps to the same (null) experience. Agents with $q = 1$ are minimally conscious because some of their cognitive states are accompanied by their single subjective experience (i.e. those in $s_1^{-1}(1)$) whereas their other cognitive states have no experience

[6] The approach essentially treats the agents of economic models as utility-maximizing algorithms, making it clear that economic modelers (at least implicitly) see consciousness as inessential for understanding behaviour.

[7] The 'Hard Problem of Consciousness', made famous by Chalmers (1996), can be thought of as understanding the microfoundations of this map.

attached (i.e. those in $s_1^{-1}(0)$).

If we let $r(\theta, \omega)$ denote the response of an agent of economic type θ to the decision problem embedded in environment ω, then "what it is like" to be a (θ, q) type agent in environment ω is given by $\ell_{\theta,q}(\omega) \equiv s_q(r(\theta, \omega))$. The function $\ell_{\theta,q}$ describes the *inner life* of a (θ, q) type agent: it reveals "what it is like" to be the agent as they traverse the various environments composing their life. Note that inner lives are expected to vary with (θ, q). They vary with θ because the map from environments to cognitive states varies with economic type (since different economic types behave differently in at least some environments), and they vary with q because the map from cognitive states to inner experiences varies with consciousness type.

Our approach to modelling consciousness has a few key features worth emphasizing. First, the notion of consciousness type allows for gradations of consciousness rather than treating it as the lights being on or off. Second, the notion of cognitive states connects consciousness with behaviour. More importantly, it emphasizes how variation in behaviour among agents suggests variation in inner experiences. To the economist's understanding that different preferences can be inferred from differences in behaviour we add that different preferences also can be inferred from differences in inner lives. Third, our approach is consistent with the view that consciousness *seems* responsible for directing behaviour even though closer examination reveals that it does not. Experience and behaviour co-move in predictable ways, but this is because of an 'omitted variable' that causes both: the cognitive state. That is, the approach is consistent with the epiphenomenalist's view that subjective experience itself has no effect on the agent's response to the decision problem that generated the experience.

But, clearly, if we want an evolutionary theory of consciousness that is built on this formalization, then the subjective experience *must* have an effect on *some* behaviour; if not in the environment responsible for the experience, then in some other. In the following section we propose that an inner life merely endows agents with information about "what it is like" to be them in various environments. This information is not inherently fitness enhancing: the value of consciousness, if any, arises *only* because individuals are social. Specifically, the value of information materializes, if at all, in the *equilibrium* of a signaling game in which reports of inner lives are (i) used to attract partners, and (ii) costly to mimic. We now turn to such a theory.

3 An Evolutionary Model of Consciousness

This section lays out a formal model with which to analyze the evolution of consciousness. We begin by describing how the decisions of agents of various types translate into fitness, and then describe how fitness differentials drive the evolution of types.

3.1 Main Ingredients

In broad strokes, the model features agents of various types that adopt personas in the attempt to find production partners. This results in fitness levels which then drives the evolution of types.

3.1.1 Agents

At any point in time, $t \in \mathbb{R}_+$, there is a continuum of agents. Each agent is endowed with a two-dimensional type, (θ, q), where $\theta \in \Theta \equiv \{0,1\}$ is the *economic type* and $q \in \{0,1,...\}$ is the *consciousness type*. Let $\Pi_t(\theta,q)$ denote the population share of (θ,q) types at date t, and let $\pi_{qt} \equiv \Pi_t(1,q)/[\Pi_t(0,q)+\Pi_t(1,q)]$ denote the proportion of agents with economic type $\theta = 1$ among those with consciousness type q at time t.

3.1.2 Production

Our model is *social* in the sense that agents produce in pairs. We model production as a Prisoners' Dilemma and interpret the economic type as an indicator for whether they are a cooperator ($\theta = 1$) or non-cooperator ($\theta = 0$). Given the agent's economic type and that of their partner, $\tilde{\theta}$, the agent's output is given by $u(\theta,\tilde{\theta})$, where

$$u(1,0) \leq u(0,0) \leq u(1,1) \leq u(0,1). \tag{1}$$

That is, an agent's output is higher when they are paired with a cooperator, but also when they themselves are a non-cooperator.

As will become clear, it is useful to define the benefit of pairing with a cooperator for type θ:

$$\delta_\theta \equiv u(\theta,1) - u(\theta,0). \tag{2}$$

The prisoners' dilemma structure implies $0 < u(1,1) - u(0,0) < \min\{\delta_0,\delta_1\}$ but it does not restrict the relative magnitudes of δ_0 and δ_1. We refer to the

case of $\delta_0 > \delta_1$ as 'decreasing differences', the case of $\delta_0 < \delta_1$ as 'increasing differences', and the case of $\delta_0 = \delta_1$ as 'equal differences'.

In order to avoid trivial cases, we assume

$$u(1,1) - u(0,0) > \psi_0. \tag{3}$$

That is, we ignore cases where the biological cost of even the most rudimentary consciousness ($q = 1$) outweighs the full gains from cooperation. Consciousness can't possibly emerge in such cases, but purely because of biological costs.

Note that, conditional on partner's economic type, an agent's consciousness type has no effect on output. If consciousness is relevant at all, it is purely because it affects who matches with whom.

3.1.3 Matching

Agents strive to pair with desirable partners by adopting a *persona*. By this we mean an agent's public projection of their inner life. This projection may be a truthful reflection of their actual inner life, or it may be an attempt to mimic the inner life of the other economic type.[8] We assume an agent's partner is randomly selected from those who share the agent's persona (or randomly selected from the population if no one shares the agent's persona).

We are agnostic as to the finer details of how mimicking is achieved and instead parameterize the cost of mimicry to capture various possibilities. An agent's consciousness type affects the cost of mimicking the other economic type. In particular, those with a higher consciousness type are more costly to mimic (owing to the greater richness of the mimicked inner life). The mimic cost is just some non-negative and increasing function of q:

$$C^{\text{Mim}}(q) \equiv \kappa_0 + \kappa \cdot c(q-1), \tag{4}$$

where c is a strictly increasing function with $c(0) = 0$, $\kappa_0 \geq 0$ is the value of $C^{\text{Mim}}(1)$ and acts as a fixed cost, whereas $\kappa \geq 0$ is a parameter that scales

[8] To simplify the exposition, we rule out the possibility of mimicking a different consciousness type. This is innocuous under the reasonable assumption that mimicking a higher consciousness type involves a mimic cost which is at least as large as the biological cost of actually possessing that higher consciousness type. In this case, agents that would mimic a different consciousness type will never survive natural selection since they must be less fit than otherwise identical agents that are actually of the mimicked consciousness type.

the marginal cost of mimicking. The critical feature of this formulation is that an agent's consciousness type q determines how costly it is for other agents to mimic them.

3.1.4 Payoffs

An agent's *economic payoff* is simply their output minus any mimic costs. To be more explicit, let $\sigma(\tilde{\theta}|\theta, q; \pi_{qt})$ denote the probability that a (θ, q)-type adopts the persona of a $(\tilde{\theta}, q)$ type when the population proportion of cooperators among those with consciousness type q is π_{qt}. The *match quality* associated with the $(\tilde{\theta}, q)$ persona is the probability that an agent with a $(\tilde{\theta}, q)$ persona pairs with a cooperator. Given σ, this match quality is therefore:

$$p(\tilde{\theta}, q; \sigma, \pi_{qt}) = \frac{\sigma(\tilde{\theta}|1, q; \pi_{qt}) \cdot \pi_{qt}}{\sigma(\tilde{\theta}|1, q; \pi_{qt}) \cdot \pi_{qt} + \sigma(\tilde{\theta}|0, q; \pi_{qt}) \cdot (1 - \pi_{qt})}. \tag{5}$$

Thus, if an agent of type (θ, q) adopts the persona of type $(\tilde{\theta}, q)$, then their *economic payoff* is their expected output net of mimic costs:

$$v(\tilde{\theta}|\theta, q; \sigma, \pi_{qt}) \equiv p(\tilde{\theta}, q; \sigma, \pi_{qt}) \cdot u(\theta, 1) + (1 - p(\tilde{\theta}, q; \sigma, \pi_{qt})) \cdot u(\theta, 0) - \mathbb{I}_{(\tilde{\theta} \neq \theta)} \cdot C^{\text{Mim}}(q), \tag{6}$$

where $\mathbb{I}_{(\tilde{\theta} \neq \theta)}$ is an indicator for whether the agent mimics. As usual, a Nash equilibrium is a profile of strategies, $\sigma^*(\cdot|\theta, q; \pi_{qt})$, such that

$$\sum_{\tilde{\theta} \in \Theta} \sigma^*(\tilde{\theta}|\theta, q; \pi_{qt}) \cdot v(\tilde{\theta}|\theta, q; \sigma^*, \pi_{qt}) \geq \sum_{\tilde{\theta} \in \Theta} \sigma(\tilde{\theta}|\theta, q; \pi_{qt}) \cdot v(\tilde{\theta}|\theta, q; \sigma^*, \pi_{qt}) \tag{7}$$

for all $\sigma(\cdot|\theta, q; \pi_{qt}) \in \Delta(\Theta)$ and all $(\theta, q; \pi_{qt})$.

Consciousness involves a biological cost, whereby higher consciousness types incur a greater cost because they produce a greater variety of experiences. This cost is normalized to zero for non-conscious agents, $q = 0$, and for $q \geq 1$ is given by

$$C^{\text{Bio}}(q) \equiv \psi_0 + \psi \cdot b(q - 1) \tag{8}$$

where b is a strictly increasing function with $b(0) = 0$, $\psi_0 > 0$ is the value of $C^{\text{Bio}}(1)$ and acts as a fixed biological cost, and $\psi > 0$ parameterizes the marginal biological cost.

The *fitness* of (θ, q)-types is their Nash equilibrium economic payoff net of the biological costs associated with consciousness:

$$f(\theta, q|\pi_{qt}) \equiv \sum_{\tilde{\theta} \in \Theta} \sigma^*(\tilde{\theta}|\theta, q; \pi_{qt}) \cdot v(\tilde{\theta}|\theta, q; \sigma^*, \pi_{qt}) - C^{\text{Bio}}(q). \tag{9}$$

3.2 Dynamics

Given the above main ingredients, we can assign agents a fitness level for any given distribution of types. We now describe how the distribution of types evolves over time as fitter types become more common.

The conditional type distribution evolves according to standard replicator dynamics, which here implies:

$$\dot{\pi}_{qt} = \left[f(1, q|\pi_{qt}) - \bar{f}(q|\pi_{qt}) \right] \cdot \pi_{qt}, \tag{10}$$

where $\bar{f}(q|\pi_{qt})$ is the average fitness of those with consciousness type q:

$$\bar{f}(q|\pi_{qt}) \equiv \pi_{qt} \cdot f(1, q|\pi_{qt}) + (1 - \pi_{qt}) \cdot f(0, q|\pi_{qt}). \tag{11}$$

The long run conditional share of cooperator types, denoted π_q^*, is a stable steady state of (10). The long run fitness associated with consciousness type q is given by:

$$f^*(q) \equiv \bar{f}(q|\pi_q^*). \tag{12}$$

Comparing long run fitness across q allows us to then derive the long run marginal distribution of consciousness types. It is natural to suppose that, in the long run, the marginal distribution of consciousness types puts all of its weight on the global maximizer of f^*. However, this approach is too permissive in the sense that it does not respect the 'uphill' nature of evolution. To illustrate, suppose it turned out that long run fitness is globally maximized at $q = 2$, yet non-conscious agents ($q = 0$) are fitter than minimally conscious agents ($q = 1$). Although $q = 2$ is the global maximizer, it is not clear how such agents arise given that $q = 1$ types are less fit than $q = 0$ types. If we start with unconscious agents ($q = 0$), then mutations to $q = 1$ arise. But this mutant population will die out if they have lower fitness than the $q = 0$ incumbents, making it implausible that $q = 2$ mutants will arise. In this case, the population gets 'stuck' at $q = 0$ rather than evolving to the global maximum. More generally, consciousness will not evolve if $f^*(0) \geq f^*(1)$, *regardless* of how f^* behaves at $q > 1$.

To capture the uphill nature of evolution, we take the long run marginal distribution of consciousness types to put all of its weight on the 'smallest local maximizer' of f^*,

$$q^* \equiv \min\{q \in \mathbb{N}_0 \mid f^*(q) \geq f^*(q+1)\}, \tag{13}$$

which we refer to as the *long run consciousness level*. We emphasize that our approach here is conservative: it can only make it more difficult for consciousness to arise since the smallest local maximizer can never be greater than the global maximizer.

We say that long run consciousness is *trivial* if $q^* \leq 1$ and *non-trivial* otherwise. We refer to $\pi^* \equiv \pi^*_{q^*}$ as *long run cooperation*, and $f^* \equiv f^*(q^*)$ as the *long run fitness*.

4 Main Results

We describe the conditions under which non-trivial consciousness may emerge in the long run. Doing so reveals a fundamental upper bound on long run consciousness, but we show that the model can generate arbitrarily large long run consciousness. To get there, we begin with some general observations.

4.1 Preliminary Observations

We begin with some preliminary observations that will help us simplify the subsequent presentation.

First, long run outcomes for non-consciousness agents–i.e. those with consciousness type $q = 0$–are straightforward: since they are unable to differentiate cooperators from non-cooperators, cooperators will be driven to extinction. Thus $\pi^*_0 = 0$ and $f^*_0 = v^*_0 = u(0,0)$. This then provides the fitness level that must be exceeded if long run consciousness of some form is to evolve.

Second, only non-cooperators mimic in equilibrium (see Lemma 2). This intuitive result allows us to focus on the incentive for non-cooperators to mimic.

Third, if, for some q, the mimic cost is higher than the benefit from pairing with a cooperator–i.e. if $C^{\text{Mim}}(q) \geq \delta_0$–then clearly no agent of consciousness type q (or higher) mimics. The long run outcomes for such agents are also straightforward: since no agent mimics, there is perfect segregation on economic type and therefore non-cooperators will be driven to extinction. Thus $C^{\text{Mim}}(q) \geq \delta_0$ implies $\pi^*_q = 0$ and $f^*_q = v^*_q = u(1,1)$.

For all other values of q, the Nash equilibrium of the mimicking game involves non-cooperators mimicking with a positive probability (Lemma 3). Equilibrium match quality lies in $[p^{\text{IM}}(q), 1]$, where $p^{\text{IM}}(q) \in (0, 1]$ is the match quality that makes non-cooperators indifferent to mimicking.

4.2 When does non-trivial consciousness emerge?

We begin by showing that non-trivial consciousness can only arise when (i) payoffs display decreasing differences, and (ii) the fixed cost of mimicry is sufficiently large, but not too large. We explain how the fixed mimicry cost cannot be too large because of a fundamental upper bound.

Proposition 1 (First Necessary Condition for Non-Trivial Consciousness)**.** *Non-trivial consciousness arises only if payoffs display decreasing differences:*

$$q^* > 1 \Rightarrow u(1,1) - u(1,0) < u(0,1) - u(0,0). \tag{14}$$

Proof: See Appendix.

Decreasing differences is necessary for consciousness because, without it, (i) the population share of cooperators among minimally conscious agents ($q = 1$) goes to either zero or one, and (ii) this implies consciousness cannot evolve further. We explain each of these in turn.

Without decreasing differences, one economic type dies out. Without decreasing differences, population dynamics are 'self-propelling' in the sense that a fitness advantage at some conditional distribution of economic types induces a change in this distribution which weakly increases the fitness advantage, thereby inducing further changes in the distribution, and so on until the conditional distribution of economic types hits a corner.

To see this formally, note that the evolution of economic types, described by (10), can be written:

$$\dot{\pi}_{qt} = \Delta(\pi_{qt}, q) \cdot \pi_{qt} \cdot (1 - \pi_{qt}). \tag{15}$$

where $\Delta(\pi_{qt}, q) \equiv f(1, q|\pi_{qt}) - f(0, q|\pi_{qt})$ is the fitness advantage of cooperator types over non-cooperative types. Letting $p^*(\pi_{qt})$ denote the Nash equilibrium match quality for agents with q-consciousness at time t, the fitness of cooperators can be written:

$$f(1, q|\pi_{qt}) = u(1,0) + p^*(\pi_{qt}) \cdot \delta_1 - C^{\text{Bio}}(q), \tag{16}$$

whereas the fact that some non-cooperators mimic in equilibrium (Lemma 3) means that the fitness of non-cooperators is the fitness of mimickers:

$$f(0,q|\pi_{qt}) = u(0,0) + p^*(\pi_{qt}) \cdot \delta_0 - C^{\text{Mim}}(q) - C^{\text{Bio}}(q). \tag{17}$$

Therefore, the difference in fitness is:

$$\Delta(\pi_{qt}, q) = p^*(\pi_{qt}) \cdot [\delta_1 - \delta_0] + C^{\text{Mim}}(q) - [u(0,0) - u(1,0)]. \tag{18}$$

The value of p^* is weakly increasing in π_{qt} (Lemma 3), and therefore the sign of the impact of π_{qt} on $\Delta(\pi_t, q)$ is the sign of $\delta_1 - \delta_0$. Without decreasing differences, the sign is non-negative and therefore $\dot{\pi}_{qt}$ is weakly increasing in π_{qt}. This implies there may be multiple stable steady states, but all of them are at the corners: $\pi^*_q \in \{0,1\}$.

Non-trivial consciousness cannot arise if one economic type dies out. If one of the economic types dies out among the minimally conscious agents, then consciousness can't evolve further. The reason is an application of the following principle.

Lemma 1 (Coexistence Principle)**.** *If $\pi^*_{\tilde{q}} \in \{0,1\}$ for some $\tilde{q} \geq 1$, then $q^* \leq \tilde{q}$.*

This principle says that the long run coexistence of economic types among $\tilde{q}$-conscious types is a necessary condition for long run consciousness to exceed $\tilde{q}$. The proof is straightforward and instructive, so we present it here. Consider some consciousness level, $\tilde{q} \geq 1$. If evolution results in all agents with consciousness $\tilde{q}$ being non-cooperators (i.e. $\pi^*_{\tilde{q}} = 0$), then their matching prospects are identical to the matching prospects of the non-consciousness agents. But then $\tilde{q}$-consciousness levels have a lower fitness than non-consciousness agents because of the biological costs associated with the consciousness agents. Thus, consciousness cannot evolve beyond $\tilde{q}$. Conversely, if evolution results in all $\tilde{q}$-consciousness agents being cooperators (i.e. $\pi^*_{\tilde{q}} = 1$), then their matching prospects are the most favourable possible. But then $\tilde{q}$-consciousness agents must have a higher fitness than agents at any higher consciousness level since they have a lower biological cost. Once again, consciousness cannot evolve beyond $\tilde{q}$.

Thus, without decreasing differences, the population evolves to a corner and this ensures trivial long run consciousness. But this need not be true with decreasing differences. In this case, the relative fitness of cooperators *decreases* as they become more common in the population. If an economic type has a fitness advantage over the other at some conditional

distribution, then the resulting evolution of the conditional distribution will act to offset that fitness advantage. This opens the possibility that the long run conditional share of cooperators among the minimally conscious is interior: $\pi_1^* \in (0,1)$. But, even in such a case, non-trivial consciousness may not arise. For example, even if minimally conscious agents enjoy a superior match quality in the long run, this additional fitness this delivers may be insufficient to offset the biological cost associated with minimal consciousness.

Proposition 2 (Second Necessary Condition for Non-Trivial Consciousness)**.** *Non-trivial consciousness arises only if the fixed mimicking cost, κ_0, is neither too large nor too small:*

$$q^* > 1 \Rightarrow \underline{C}' < \kappa_0 < \overline{C}, \tag{19}$$

where $\underline{C} \equiv (u(0,0) - u(1,0)) \cdot \frac{\delta_0}{\delta_1}$, $\underline{C}' \equiv \underline{C} + \frac{\delta_0 - \delta_1}{\delta_1} \cdot \psi_0$, *and* $\overline{C} \equiv u(0,1) - u(1,1)$.

Proof: See Appendix.

The level of the mimic cost for minimally conscious agents, κ_0, is relevant because it affects the long run proportion of cooperators among them, π_1^*. In particular, equation (18) shows that higher values of κ_0 raises $\Delta(\pi_{qt}, q)$ and, therefore, the proportion of cooperators among minimally conscious agents, π_1^*. This, in turn, raises the long run fitness of minimally conscious agents, f_1^*, since such agents more readily match with cooperators. In short, the long run fitness of minimally conscious agents is increasing in κ_0.

Mimic costs can't be too small. If the mimic cost is too small–$\kappa_0 \leq \underline{C}'$–then the long run fitness of minimally conscious agents is too low. Specifically, it is lower than the fitness of non-conscious agents and therefore non-trivial consciousness can't arise. If mimic costs are particularly small–$\kappa_0 \leq \underline{C}$–then minimally conscious cooperators actually die out. This is a "Green Beard" situation: any period in which cooperators are able to profitably encounter each other is short-lived as non-cooperators move in. Cooperators will not die out for $\kappa_0 \in (\underline{C}, \underline{C}')$, yet will survive in a small enough proportion that the minimally conscious agents have a lower fitness than non-conscious agents (because of the biological costs).

Mimic costs can't be too large. On the other hand, if the mimic cost is too large–$\kappa_0 \geq \overline{C}$–then non-trivial consciousness cannot evolve because cooperators come to dominate among the minimally conscious ($\pi_1^* = 1$). The

coexistence principle then rules out the possibility of consciousness evolving further. The long run extinction of non-cooperators among the minimally conscious is not necessarily because κ_0 is so high as to kill off the incentive to mimic, condemning non-cooperators to pair with other non-cooperators. Rather, minimally conscious non-cooperators die out because the high mimic cost implies a fitness disadvantage so large that it persists even as they approach extinction.

4.3 Human Level Consciousness?

The argument that mimic costs cannot be too large suggests a larger point: long run consciousness is fundamentally bounded, in the sense that q^* will be finite even in the absence of biological costs. In short, the coexistence principle says that consciousness cannot evolve beyond the point at which mimic costs are sufficiently large to cause non-cooperators to die out. But such a point will eventually be reached at some consciousness level. In other words, the fundamental upper bound on long run consciousness, denoted $\bar{q}$, is given by the smallest conscious level with an associated mimic cost that exceeds $\overline{C}$:

$$q^* \leq \bar{q} \equiv \min\{q \mid C^{\text{Mim}}(q) \geq \overline{C}\}. \tag{20}$$

Note that this upper bound is decreasing in the marginal mimic cost parameter, κ.

We have seen that many parameter configurations lead to evolutionary paths that terminate in trivial consciousness and therefore look nothing like human level consciousness. Even if non-trivial consciousness does evolve, the existence of a fundamental upper bound suggests there is no guarantee that it will be large enough to plausibly resemble human-level consciousness. For instance, the bound implies that long run consciousness can't be made arbitrarily large by taking the marginal biological cost, ψ, to zero. Similarly, taking the marginal mimic cost, κ, to zero will make the upper bound arbitrarily large but it will not lead to arbitrarily large consciousness. Rather, it will lead to *trivial* consciousness. This is because total mimic costs will become increasingly similar for all $q \geq 1$, implying that agents get the same long run economic payoff for all $q \geq 1$. But biological costs are lowest, and thus fitness the highest, at $q = 1$.[9]

Given the above observations, an important question is whether the model can generate long run consciousness large enough to resemble the

[9] In Lemma 7 we show that long run consciousness is weakly decreasing in ψ and is hump-shaped in κ.

richness of human consciousness. While we do not have strong priors as to the complexity of human level consciousness, our next result shows this does not matter because the model can generate arbitrarily large long run consciousness if parameters are appropriately chosen.

Proposition 3 (Consciousness Can Be Made Arbitrarily Large)**.** *Assume decreasing differences. For any $\tilde{q}$, there exists positive parameters $(\{\kappa_0, \kappa\}, \{\psi_0, \psi\})$ such that $q^* \geq \tilde{q}$.*

Proof: See appendix.

Proposition 3 indicates that, starting from a population of non-conscious agents, any level of conscious complexity one supposes to correspond to human level consciousness, $\tilde{q}$, will be reached by evolution for some parameter configurations of this model. The fixed cost parameters (κ_0, ψ_0) are chosen so that the necessary condition from Proposition 2 is satisfied. The value of κ is then chosen small enough that the fundamental upper bound exceeds $\tilde{q}$. Since $\kappa > 0$, the mimic cost strictly increases with q. Since $\tilde{q} < \bar{q}$, the long run economic payoff is strictly increasing in q (up to at least $\tilde{q}$) as increasingly costly mimicking leads to increasingly superior long run match qualities. The value of ψ is then chosen small enough that biological costs are sufficiently negligible that long run fitness at $\tilde{q}$ is higher than at any lower value of q.

Finally note that, apart from decreasing differences, the result is independent of the parameters of the prisoner's dilemma game. Specifically, there are parameters under which any level of consciousness can be attained irrespective of the gains from cooperation relative to defection and the distance between cooperation and equilibrium payoffs in the prisoner's dilemma game.

4.4 Characterizing Long Run Consciousness

We now move from necessary conditions to a characterization. This will allow us broader insight into the conditions under which non-trivial consciousness will emerge.

Proposition 4. *Under decreasing differences, the difference in long run economic payoff between q-consciousness types and non-conscious types is:*

$$v_q^* - v_0^* = \Phi(q) \equiv \max\{0, \min\{\phi(q), \Delta\}\}, \tag{21}$$

where $\Delta \equiv u(1,1) - u(0,0)$ *is the gains from cooperation and*

$$\phi(q) \equiv \frac{\delta_1}{\delta_0 - \delta_1} \cdot \left[C^{Mim}(q) - \underline{C} \right]. \tag{22}$$

Proof: See appendix.

Since $f_q^* - f_1^* = v_q^* - v_1^* - C^{\text{Bio}}(q)$, we have that q^* is the smallest local maximizer of:

$$F(q) \equiv \Phi(q) - C^{\text{Bio}}(q). \tag{23}$$

The necessary conditions on the mimic cost are apparent. Since $F(0) = 0$, the smallest local maximizer is $q^* = 0$ if $F(1) \leq 0$ (this is equivalent to $C^{\text{Mim}}(q) \leq \underline{C}'$). Similarly, the smallest local maximizer is $q^* = 1$ if $\Phi(1) = \Delta$ (this is equivalent to $C^{\text{Mim}}(q) > \overline{C}$). The fundamental upper bound is the smallest q for which $\Phi(q) = \Delta$.

But the characterization is useful for evaluating the effect of parameters. For instance, suppose we parameterize payoffs letting $\alpha_1 \equiv u(0,1) - u(1,1)$, $\Delta \equiv u(1,1) - u(0,0)$, and $\alpha_0 \equiv u(0,0) - u(1,0)$. Each of these are nonnegative and decreasing differences amounts to $\alpha_1 > \alpha_0$. We can show that long run consciousness is weakly increasing in the gains from cooperation, Δ, since this raises the level and the slope of Φ. Yet, the fundamental upper bound is unaffected by Δ and therefore q^* will not become arbitrarily large.

4.5 How are consciousness and cooperation related?

The model describes the coevolution of consciousness type and economic type and thus says something about the relationship between consciousness and cooperation.

Biologically costly consciousness survives evolutionary selection because it helps cooperators match better and reap mutual fitness gains. Consciousness and cooperation are thus symbiotically connected in our setting. This allows us to derive a necessary condition for the existence of consciousness.

Corollary 1 (Economic Correlates of Consciousness)**.** *A population's constituents are conscious only if the population has a positive proportion of cooperator types.*

The model features an "if and only if" relationship between consciousness and cooperation. But the sufficiency is an artifact of our framework being devoid of the other mechanisms by which (non-payoff maximizing)

cooperators could survive fitness based selection. So it does not follow that the presence of non-kin based cooperation necessarily implies the presence of consciousness. But the prediction that absent non-kin cooperation consciousness cannot evolve would seem to rule out consciousness in many species.

5 Discussion

5.1 How does our explanation avoid our critique of existing theories?

We have noted existing theories of consciousness largely amount to describing some mental activity purportedly associated with consciousness and then explaining the function of that activity without explaining why consciousness *per se* is needed. We now elaborate on how our model avoids this criticism.

5.1.1 Why the Properties of Consciousness?

The literature on consciousness within philosophy has asked the following question. "What is the (or a) function that consciousness and consciousness alone could perform?". The answer, which Chalmers (1995) and many others have argued, is none. For any possible function put forward as requiring consciousness, one counter-posits an alternative, more parsimonious, mechanism that bypasses the experiential component yet yields the same function. For example, one may posit that consciousness is required to deliberate on a plan of action. In principle a plan can be improved by subjecting it to scrutiny and deliberation. This begs the question of why scrutiny and deliberation require consciousness? What is it about these that require subjective first-person experiences? Whatever mental modules are applied to improve a plan, there seems nothing functionally gained by adding an experiential accompaniment to these modules. Similar reasoning can be applied to rule out consciousness arising due to any other posited functions; see Chalmers (2010) p. 16.

Here, we use the model to proceed in a different direction. Instead of asking what potential useful functions require a subjective experiential accompaniment? The answer to which seems to be nothing. We instead ask: "What properties would a function that helps cooperators match by sharing subjective experienced content have to have in order for that function to be selected by evolution?" According to our theory, specific properties

are likely and, in some cases, essential for such a function to survive selection. We show that key features of consciousness are thus predicted by our theory, and others, though not predicted, can be understood as likely to arise under evolutionary pressure.

Privacy Inner experiences are private; see Léon (1997). Could evolution have alternatively selected observable inner experiences? The answer is no if it evolved to facilitate cooperator matching.

If agents' inner experiences were observable, perhaps imperfectly as in Frank (1987), then there would be no payoff benefit to the experiential component; and costly inner experiences could not survive selection pressure. If agents could directly observe the inner experiences of each other they would not need to rely on messages of such experience to match, but would know them immediately, and thereby know type as well. With type known, there would no longer be any additional fitness advantage to generating subjective first-person experiences. Whatever could be messaged to a potential partner would already be known, and mimicry would not be possible. So any metabolic costs borne of generating experience would be costs without fitness benefit, and would be selected against. Cooperation would emerge directly if type were observable, because cooperators would be known and directly rewarded. But the experiential component – that is consciousness itself – would be redundant.

So, according to our model, if subjective inner experiences evolved to facilitate cooperator matching, then these must necessarily be private in nature. Publicly observable subjective experiences could not evolve.

Type Specificity In principle, it is conceivable that we may have evolved inner experiences that bear no correspondence with economic type. Yet, for evolution to select experience for the reasons posited in our model this would again not be possible. Agents of different types must have distinct experiential labels over at least some common cognitive states. For example, a cooperator type has an inner experience of 'cheating' that is systematically distinct from a non-cooperator. This makes projecting an inauthentic persona (learning the inner experiences of another type so as to message consistently about them) costly (and equally so for all types).[10] Our model therefore predicts that, necessarily, subjective inner experiences vary by economic type (as there is evidence that they do in reality; see appendix section C.2).

[10] In contrast to, for example, Kartik (2009) who assumes a distaste for deception.

Seriality Why are humans only able to be consciously aware of a single percept at once? Though possible that this simply results from a hard physiological constraint that is itself biologically costly to overcome, this would seem unlikely since the brain is highly parallel in most other cognitive processes, and able to undertake many complex tasks of homeostasis, for example, and non-conscious processing simultaneously. In contrast, humans are strangely serial in their stream of consciousness. Many experiments have shown that we cannot simultaneously be conscious of separate objects presented to different parts of (for example) the visual field. Conscious attention flips from one object to the other in sequence, suggesting the presence of a bottleneck in conscious attention; see Dehaene (2014) pp. 27-30, for a discussion. Why didn't nature select for the possibility of simultaneous conscious streams?

Unlike the two features above, our model does not strictly predict that seriality is the only form of inner experience that could evolve, but it does suggest an explanation for why it may be so. Since conscious experiences arise to inform messages regarding type and enable matching, the existence of other bottlenecks along the line of communication would imply that bearing the metabolic cost of relieving them at the point of consciousness would create little selective benefit. Given the human messaging system only affords the possibility of serial communication, (verbal messaging is constrained by, amongst other things, humans having only a single mouth and set of vocal chords) there would seem to be little fitness gain from having parallel experience streams to create conscious content. Having multiple streams would allow faster processing of, for example, imagined subjective experiences and create the content required more quickly. But this content would still meet the seriality constraint arising with communication. We can only tell others of our inner experiences one experience at a time, because we can only communicate one percept or concept at a time. So even if feasible at some metabolic cost, the addition of seriality to the human hardware of consciousness would create no fitness gain given the other constraints on communication that humans have. Such a capacity would be unlikely to survive selection.

Ineffability As has long been commented upon, experienced inner states have an ineffable quality. The experience of, for example, tasting chocolate, or seeing a colour, cannot be perfectly described in language and communicated so that listeners can unambiguously understand the subjective experience of another.

> The arguably most crucial features attributed to phenomenal experience are its essential subjective nature, which sometimes is taken to mean that phenomenal consciousness embodies a particular point of view, but also that some of its parts or properties seem ineffable, private, or unavailable to cognitive and linguistic processing or communication. Kleiner (2020)

In principle, there is no reason that subjective, first-person experiences could not be fully articulate, in the same way that mathematical objects can be completely described. Could we have evolved subjective experiences like that instead? Once again, the model does not strictly predict ineffability, but does suggest why it may have this feature.

Suppose that articulate inner lives are available at some biological cost. For simplicity suppose that a highly unrealistic "articulate mutation" exists. That is, one which would allow recipients to fully articulate and fully comprehend the articulations of others' inner experiences. We can use the model to ask whether this characteristic is likely to be valuable in matching? To contemplate this, consider an example where non-cooperators and cooperators have the following distinct subjective first-person experiences and also assume an extremely, and unrealistically simple, form of messaging distinguishes type. A cooperator type truthfully reports the following phenomenological experiences " I see the colour red when somebody cheats me, and I similarly see the colour red when someone else is cheated." A mimicker sees the colour red when somebody cheats them, but not when they see someone else cheated. According to the model this gives cooperators an advantage in matching which could be overcome by the mimicker, but only at the cost of acquiring the information regarding what a cooperator type subjectively "sees" when someone else is cheated. But note that the efficacy of cooperator matching, and corresponding difficulty of non-cooperators mimicking, is unaffected by the ineffability of 'red'. That is, even though it is impossible for an agent to know that what they experience as red is the same as what another agent reports as red, this does not matter. All that matters is that the subjective experience 'see red' accompanies two distinct cognitive states, and what needs to be understood is that these cognitive states are reported as generating an equivalent subjective experience. Adding the capacity for 'see red' to be fully articulable, so that one agent knew precisely what it meant for another to see red, and could be sure that they had experienced the same subjective phenomenon as the other, would not help in matching. The biological cost associated with articulate instead of ineffable phenomenology would not be rewarded with a

fitness payoff, so any cost to adding the articulate capacity would seem to be a cost borne without a corresponding fitness payoff.

As the example makes clear, articulate subjective experiences are by no means necessary to the function of enabling matching via establishing distinct experiences. So, though ineffable phenomenology is not a strong prediction of the model, as the previous features, privacy, type specificity and seriality were, it is certainly consistent with the model. And if this function is costly, the model suggests reasons why it would not have been selected during our evolution, and why we are endowed with only ineffable subjective experiences.

Imagined Subjective Experiences Humans are able to imagine the nature of subjective experiences they will have in entirely novel, and even never-to-be-experienced situations. Our model suggests why this capacity to imagine subjective experiences may be evolutionarily helpful. Subjective first-person experiences that accompany particular objective or cognitive states are used by agents to inform the statements (how they feel) that will allow them to match and separate. But the set of potential external environments, and hence cognitive states, that humans traverse is enormous. If generating inner experiences across this set required actually being in each such cognitive state, then for an agent to know their own subjective experiences would require a large resource investment. A cost saving innovation would be the ability to concoct and discuss entirely hypothetical environments, and thereby generate both the cognitive states and their accompanying subjective inner experiences, via simulation. Being able to imagine how one would feel (one's subjective experiences) in never experienced situations, allows agents to do just this. It allows them to furnish, and assess, detailed descriptions of their own, and others', subjective experiences without having to bear the expense of exploring such cognitive states themselves. By allowing agents to compare subjective first-person inner experiences over a much broader range of situations than can actually be experienced, it helps agents obtain the content required to compare their cognitive state labels and to match by type. The model thus suggests reasons for why even metabolically costly 'imagined subjective experiences' would have survived evolutionary selection amongst humans.

So, to conclude this section, for a human facility to have the function of helping cooperators match by sharing subjectively experienced content, our evolutionary model predicts that such a facility must exhibit both privacy and type-specificity as human consciousness does. Additionally,

though not strictly predicted, the model provides reasons for why this facility would be serial only, why it would remain ineffable, and why the capacity to imagine subjective experiences may survive selection.

5.2 What does all this say about consciousness in non-human systems?

Now that we have a theory of how consciousness emerges, we can apply the theory to the question of where else consciousness is likely to exist.

Consciousness requires social interaction of a particular form: a prisoners' dilemma. Moreover, a particular case of the prisoners' dilemma in which an increase in the probability of pairing with a cooperator benefits a non-cooperator more than a cooperator (decreasing differences).

It also requires hidden types and costly mimicry.

5.2.1 Required Social Conditions

The structure of social interactions required for consciousness to emerge is non-trivial. Here we have treated this structure as a primitive – already in place – though it would seem that a more sophisticated model could have it co-evolve with the structure in place here.

For example, communication is necessary to convey a common understanding of external events. Agents must be able to communicate that they are in similar cognitive states so that they can compare respective inner experiences. For humans, these descriptions can be extremely complicated because the vast number of possible inner states triggered by distinct inner experiences is huge. They need not be for lower q species. It would seem however that consciousness, in addition to being restricted to 'social' species, would further seem to require some level of linguistic (or other form of messaging) sophistication. We have not explored how the development of consciousness may also affect the complexity of the messaging system, and how they may interact and co-evolve. This seems an interesting avenue to consider further.

6 Conclusions

The evolutionary analysis we undertake yields a very different explanation (and predictions) than the non-evolutionary theories that predominate within consciousness studies. If we are accurate, then the value of consciousness has nothing to do with what is being processed in the brain

(unlike Global Workspace Theories such as Dehaene (2014)), nor with how it is being processed in the brain (unlike Integrated Information Theories such as Tononi et al. (2016)). It emerged in social species only, as a by-product of an evolutionary struggle to attract desirable partners. A prediction of the framework is that, necessarily, any species that has developed consciousness must also have developed a substantial proportion of agents who act like these "desirable partners", i.e. who undertake costly (fitness reducing) actions that benefit their non-kin partners. Consciousness would thus seem to require the existence of agents with a tendency to act selflessly, prosocially or altruistically.

Appendix

A Proofs and Further Results

As mentioned in the text, outcomes for non-consciousness agents (i.e. those with consciousness type $q = 0$) levels are straightforward: since they are unable to differentiate cooperators from non-cooperators, cooperators will be driven to extinction. Thus $\pi_0^* = 0$ and $f_0^* = v_0^* = u(0,0)$. The rest of the analysis presented here derives outcomes for $q \geq 1$.

Our approach to analysing the model involves three stages. The first derives equilibrium mimicking, taking the consciousness type and conditional distribution of economic types as given. The second involves endogenizing the conditional distribution by deriving its evolutionary dynamics. This step also produces a long-run fitness level associated with each q-consciousness level. The third stage derives the long run consciousness level by comparing long run fitness across each q.

A.1 Stage 1: Static Mimicking Equilibrium

In this section we derive Bayesian Nash equilibria of the mimic game given an arbitrary consciousness type and conditional distribution. The value of q affects both mimic and biological costs but only the mimic costs affect incentives to mimic. As such, we temporarily treat the mimic cost as a parameter, denoted C^{Mim}, dropping an explicit reference to q to simplify notation. Similarly, and again temporarily, the share of cooperators (π_{qt}) is denoted simply as π.

We begin with an intuitive result that allows us to simplify the analysis to follow.

Lemma 2. *In any Bayesian Nash Equilibrium of the mimic game, all cooperators mimic with probability zero.*

Proof. Suppose not. Then the match success associated with a non-cooperator persona must be strictly higher than the match success associated with a cooperator persona (to cover the cooperator's mimic cost). But this implies that non-cooperators will strictly prefer to not mimic. But then the match quality associated with the cooperator persona cannot be lower than the match quality associated with the non-cooperator persona since adopting the cooperator persona ensures a match with a cooperator. This contradicts the requirement that the cooperator persona has a strictly higher match

quality. □

Given this, let σ denote the probability that a non-cooperator mimics and let p be the match quality associated with the cooperator persona. Non-cooperators are willing to mimic if the gain, $p \cdot \delta_0$, exceeds the cost, C^{Mim}. Let

$$p^{\text{IM}} \equiv \frac{C^{\text{Mim}}}{\delta_0} \tag{24}$$

denote the value of p that makes non-cooperators indifferent to mimicking ("IM" = indifferent to mimicking), and let $\hat{\sigma}(\pi)$ be the value of σ that satisfies $\pi/[\pi + (1-\pi)\cdot\sigma] = p^{\text{IM}}$. That is:

$$\hat{\sigma}(\pi) \equiv \frac{\pi}{1-\pi} \cdot \frac{1-p^{\text{IM}}}{p^{\text{IM}}}. \tag{25}$$

Given these definitions, the static equilibrium outcomes are summarized as follows.

Lemma 3. *For any π, there exists a unique static equilibrium. This equilibrium involves non-cooperators mimicking with probability*

$$\sigma^*(\pi) = \max\{\min\{\hat{\sigma}(\pi), 1\}, 0\}, \tag{26}$$

so that match success is given by

$$p^*(\pi) \equiv p(\sigma^*(\pi), \pi) = \max\{\min\{p^{IM}, 1\}, \pi\}, \tag{27}$$

where p^{IM} and $\hat{\sigma}(\pi)$ are defined in (24) and (25) respectively.

Proof. For an arbitrary matching success, p, the expected economic payoff for cooperator types and for mimicking non-cooperator types, respectively, are:

$$v(1|p) = p \cdot u(1,1) + (1-p) \cdot u(1,0) \tag{28}$$

$$v(0|p, C^{\text{Mim}}) = p \cdot u(0,1) + (1-p) \cdot u(0,0) - C^{\text{Mim}}. \tag{29}$$

Non-cooperators that do not mimic are certain to pair with a non-cooperator and thus get an economic payoff of $u(0,0)$. Non-cooperators are therefore indifferent to mimicking when p equals p^{IM} as defined in (24). In particular, non-cooperators will strictly prefer to mimic if $p > p^{\text{IM}}$ and strictly prefer to not mimic if $p < p^{\text{IM}}$. From (5) we have that p is determined by

$$p(\sigma, \pi) = \frac{\pi}{\pi + \sigma \cdot (1-\pi)}. \tag{30}$$

Thus, the Nash equilibrium probability of mimicking, σ^*, is a value satisfying:

$$\sigma^*(\pi)\begin{cases} = 0 & \text{if } \frac{\pi}{\pi+\sigma^*(\pi)\cdot(1-\pi)} < p^{\text{IM}} \\ \in [0,1] & \text{if } \frac{\pi}{\pi+\sigma^*(\pi)\cdot(1-\pi)} = p^{\text{IM}}(q) \\ = 1 & \text{if } \frac{\pi}{\pi+\sigma^*(\pi)\cdot\pi} > p^{\text{IM}} \end{cases} \tag{31}$$

All possibilities fall into one of the following three cases.

1. $p^{\text{IM}} \leq \pi$. Here each non-cooperator finds mimicking optimal even if all other non-cooperators mimic. The Nash equilibrium therefore involves all non-cooperators mimicking; $\sigma^*(\pi) = 1$ and therefore $p^*(\pi) \equiv p(\sigma^*(\pi), \pi) = \pi$.

2. $p^{\text{IM}} \in (\pi, 1)$. Here each non-cooperator strictly wants to mimic if no other non-cooperator mimics (since $p^{\text{IM}} < 1$), and each non-cooperator strictly wants to refrain from mimicking if all other non-cooperators mimic (since $\pi < p^{\text{IM}}$). Since $p(\sigma, \pi)$ falls with the mimicking frequency σ, the Nash equilibrium involves an interior proportion of non-cooperators mimicking. Specifically, σ^* is the value that satisfies $p(\sigma^*, \pi) = p^{\text{IM}}$. That is, $\sigma^*(\pi)$ equals $\hat{\sigma}(\pi)$ as defined in (25). Therefore $p^*(\pi) \equiv p(\sigma^*(\pi), \pi) = p^{\text{IM}}$ in this case.

3. $1 \leq p^{\text{IM}}$. Here each non-cooperator strictly prefers to refrain from mimicking, even if all other non-cooperators similarly refrain from mimicking. The Nash equilibrium therefore involves no mimicking; $\sigma^*(\pi) = 0$ and therefore $p^*(\pi) \equiv p(\sigma^*(\pi), \pi) = 1$.

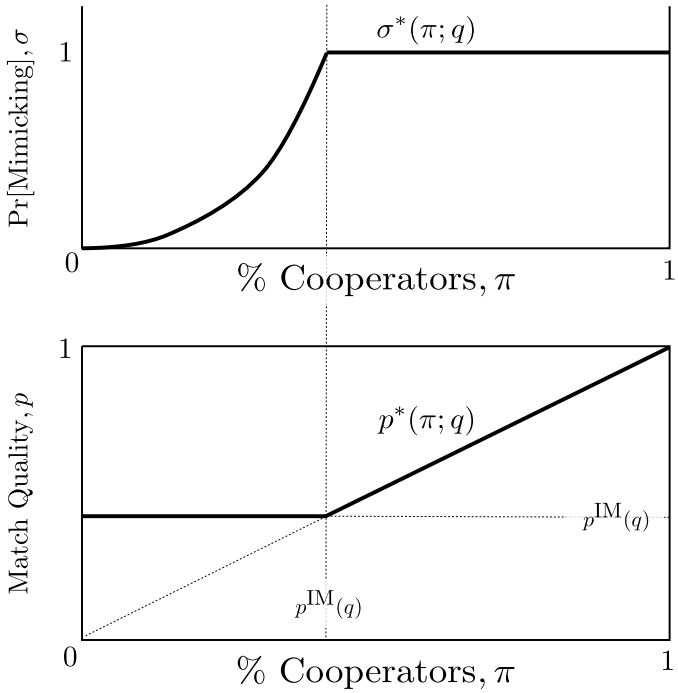
Therefore we have:

$$(\sigma^*(\pi), p^*(\pi)) = \begin{cases} (0,1) & \text{if } 1 \leq p^{\text{IM}} \\ (\hat{\sigma}(\pi), p^{\text{IM}}) & \text{if } \pi \leq p^{\text{IM}} < 1 \\ (1, \pi) & \text{if } p^{\text{IM}} \leq \pi. \end{cases} \tag{32}$$

If $p^{\text{IM}} < 1$, then $\hat{\pi}(\pi)$ increases from zero to one as π increases from zero to p^{IM}. Thus, the above gives $\sigma^*(\pi) = \min\{\hat{\sigma}(\pi), 1\}$. Alternatively, if $p^{\text{IM}} \geq 1$, then we have $\hat{\pi}(\pi) < 0$ and $\sigma^*(\pi) = 0$. Together then we have $\sigma^*(\pi)$ equals $\hat{\sigma}(\pi)$ but capped between zero and one. That is, $\sigma^*(\pi) = \max\{\min\{\hat{\sigma}(\pi), 1\}, 0\}$ as claimed.

Similarly, if $p^{\text{IM}} < 1$, then the above gives $p^*(\pi) = \min\{p^{\text{IM}}, \pi\}$. Alternatively, if $p^{\text{IM}} \geq 1$ we have $p^*(\pi) = 1$. Together then we have $p^*(\pi)$ equals π

Figure 1: Static Mimicking Equilibrium



Notes. The top panel shows the equilibrium mimicking probability for each value of π in the case of $P^{\mathrm{IM}}(q) \leq 1$. The bottom panel shows the associated match quality.

but capped below both one and p^{IM}. That is, $p^*(\pi) = \max\{\min\{p^{\mathrm{IM}}, 1\}, \pi\}$ as claimed. □

Intuitively, uniqueness arises because mimicry involves strategic substitutes: non-cooperators are more willing to mimic when fewer others are doing so. This is a straightforward consequence of more mimicry reducing the probability that a mimic ends up paired with a cooperator. In short, equilibrium involves non-cooperators mimicking with a probability that leaves non-cooperators indifferent to mimicking when possible. That is, the mimicking probability is constrained to lie in the unit interval. Similarly, matching success is that level that leaves non-cooperators indifferent to mimicking when possible. That is, matching success is constrained to lie above π (when all non-cooperators mimick) and below unity (when no one mimicks). These relationships are illustrated in Figure 1.

A.2 Evolution within q

We now endogenize the conditional distribution of economic types by allowing evolution to select economic types according to equilibrium fitness. That is, using the same notation as in the previous section, we derive π^* which allows us to derive matching success in the long run, $p^* \equiv p(\sigma^*(\pi^*), \pi^*)$. We first dispense with the trivial case where $p^{\text{IM}} \geq 1$.

Lemma 4. *If $p^{IM} \geq 1$, then $\pi \to \pi^* = 1$*

Proof. $p^{\text{IM}} \geq 1$ means that non-cooperators face a mimic cost that outweighs the maximal possible benefit of being guaranteed a cooperator partner. In such cases, cooperators will pair with cooperators and therefore have an economic payoff of $u(1,1)$, whereas non-cooperators will pair with non-cooperators and therefore have an economic payoff of $u(0,0)$ which is less than $u(1,1)$. Thus, cooperators will have a higher fitness than non-cooperators and $\pi \to \pi^* = 1$. □

We now consider the more interesting cases where $p^{\text{IM}} < 1$. Notice, from Lemma 3, that this implies

$$p(\pi) = \max\{p^{\text{IM}}, \pi\} \tag{33}$$

Since $p(\pi) \geq p^{\text{IM}}$ for all π, non-cooperators weakly prefer to mimic in equilibrium. Therefore all non-cooperators will get a payoff equal to the payoff of mimickers. Therefore, given an arbitrary $p \in [p^{\text{IM}}, 1]$, the difference in expected payoff between cooperators and non-cooperators, from (28) and (29) is:

$$v(1|p) - v(0|p, C^{\text{Mim}}) = C^{\text{Mim}} - (u(0,0) - u(1,0)) - p \cdot [\delta_0 - \delta_1]. \tag{34}$$

Therefore, cooperators and non-cooperators will get equal payoffs ("EP") when p takes the value defined by

$$p^{\text{EP}} \equiv \frac{C^{\text{Mim}} - (u(0,0) - u(1,0))}{\delta_0 - \delta_1} \tag{35}$$

Away from p^{EP}, the sign of the difference in expected payoffs depends on the sign of $\delta_0 - \delta_1$.

We know from the static analysis above that an increase in π weakly increases p, but the effect of p on the relative fitness of cooperators and non-cooperators depends on whether payoffs display increasing differences

(super-modularity), equal, or decreasing differences (sub-modularity). It turns out that the long run conditional distribution is easily characterized under increasing or equal differences.

Lemma 5. *Increasing or equal differences (generically) imply* $\pi^* \in \{0,1\}$.

Proof. Lemma 4 establishes that $\pi^* \in \{0,1\}$ when $p^{\text{IM}} \geq 1$. So here we consider the case of $p^{\text{IM}} < 1$.

To describe the long run outcomes, it is useful to define two critical values. First, let $\underline{C}$ denote the value of mimic costs that equate p^{IM} and p^{EP}:

$$\underline{C} \equiv \frac{(u(0,0)-u(1,0))\cdot(u(0,1)-u(0,0))}{(u(1,1)-u(1,0))} = (u(0,0)-u(1,0))\cdot\frac{\delta_0}{\delta_1}. \tag{36}$$

Second, let $\overline{C}$ denote the value of mimic costs that equate p^{EP} and unity:

$$\overline{C} \equiv u(0,1)-u(1,1). \tag{37}$$

We start by considering the case of $\delta_0 < \delta_1$ (i.e. increasing differences). We see from (34) that cooperators get a strictly higher payoff than mimicking non-cooperators if and only if

$$p(\pi) > p^{\text{EP}}. \tag{38}$$

Using (33), the dynamics of $\pi > 0$ are given by:

$$\dot{\pi}(\pi)\begin{cases} < 0 & \text{if } \max\{p^{\text{IM}},\pi\} < p^{\text{EP}} \\ = 0 & \text{if } \max\{p^{\text{IM}},\pi\} = p^{\text{EP}} \\ > 0 & \text{if } \max\{p^{\text{IM}},\pi\} > p^{\text{EP}} \end{cases} \tag{39}$$

If $p^{\text{IM}} \neq p^{\text{EP}}$, then the only steady states of this system are $\pi^* = 0$ or $\pi^* = 1$. Specifically, if $p^{\text{IM}} > p^{\text{EP}}$ then $\dot{\pi}(\pi) > 0$ for all π which implies $\pi^* = 1$. On the other hand, if $p^{\text{IM}} < p^{\text{EP}}$ then $\dot{\pi}(\pi) < 0$ for all $\pi \in [0, p^{\text{EP}})$ and $\dot{\pi}(\pi) > 0$ for all $\pi \in (p^{\text{EP}}, 1]$. There is a steady state at $\pi = p^{\text{EP}}$, but it is unstable. Therefore, the only stable steady states when $p^{\text{IM}} < p^{\text{EP}}$ are $\pi^* = 0$ and $\pi^* = 1$. Since $p^{\text{IM}} \neq p^{\text{EP}}$ holds except when C^{Mim} happens to equal exactly $\underline{C}$ (defined in (37)), it holds generically. Therefore it is generically true that $\pi^* \in \{0,1\}$ under increasing differences.

We now turn to the case of $\delta_0 = \delta_1$ (i.e. equal differences). We see from (34) that relative fitness does not depend on π. Using (33), the dynamics of $\pi > 0$ are given by:

$$\dot{\pi}(\pi)\begin{cases} < 0 & \text{if } C^{\text{Mim}} < \underline{C} \\ = 0 & \text{if } C^{\text{Mim}} = \underline{C} \\ > 0 & \text{if } C^{\text{Mim}} > \underline{C} \end{cases} \tag{40}$$

where $\underline{C}$ is defined in (37). We see from these dynamics that $\pi^* \in \{0,1\}$ except for the special case where $C^{\text{Mim}} = \underline{C}$. Thus, under equal differences it is also generically true that $\pi^* \in \{0,1\}$. □

Lemma 6. *Under decreasing differences, the long run conditional distribution is generically unique. The long run proportion of cooperators, π^*, matching success, p^*, and economic payoffs, v^*, are weakly increasing in mimic costs.*[11] *Specifically:*

$$[\pi^*, p^*, v^*] = \begin{cases} [0, p^{IM}, u(0,0)] & \text{if } C^{Mim} < \underline{C} \\ [p^{EQ}, p^{EQ}, u(1,0) + p^{EQ} \cdot \delta_1] & \text{if } C^{Mim} \in \left(\underline{C}, \overline{C}\right) \\ [1, 1, u(1,1)] & \text{if } C^{Mim} \geq \overline{C}. \end{cases} \tag{41}$$

Proof. Consider the case of $\delta_1 < \delta_0$ (decreasing differences), noting that this implies $\underline{C} < \overline{C}$.

We first we consider the case of $p^{\text{IM}} < 1$. From (34), cooperators get a strictly higher payoff than non-cooperators if and only if p is sufficiently *small* that $p < p^{\text{EP}}$. Using (33), the dynamics of $\pi > 0$ are given by:

$$\dot{\pi}(\pi) \begin{cases} > 0 & \text{if } \max\{p^{\text{IM}}, \pi\} < p^{\text{EP}} \\ = 0 & \text{if } \max\{p^{\text{IM}}, \pi\} = p^{\text{EP}} \\ < 0 & \text{if } \max\{p^{\text{IM}}, \pi\} > p^{\text{EP}} \end{cases} \tag{42}$$

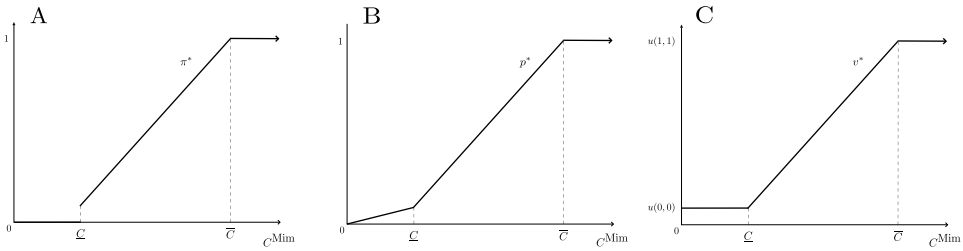
Given $p^{\text{IM}} < 1$, there are (generically) three possible cases:

1. $p^{\text{EP}} < p^{\text{IM}}$. This is equivalent to $C^{\text{Mim}} < \underline{C}$. From (42), we have $\dot{\pi}(\pi) < 0$ for all π, so that $\pi \to 0$ while $p \to p^* \equiv p(\pi^*) = p^{\text{IM}}$. The value of v^* is $u(0,0)$ since p^* leaves non-cooperators indifferent to mimicking.

2. $p^{\text{IM}} < p^{\text{EP}} < 1$. This is equivalent to $C^{\text{Mim}} \in \left(\underline{C}, \overline{C}\right)$. From (42), we have $\dot{\pi}(\pi) > 0$ for all $\pi \in [0, p^{\text{EP}})$ and $\dot{\pi}(\pi) < 0$ for all $\pi \in (p^{\text{EP}}, 1]$, so that $\pi \to \pi^* = p^{\text{EP}}$ and $p \to p^* = p^{\text{EP}}$. The value of v^* is the payoff achieved by cooperators at p^* (which will equal the payoff of non-cooperator mimickers). That is, $v^* = p^* \cdot u(1,1) + (1-p^*) \cdot u(1,0) = u(1,0) + p^{\text{EP}} \cdot \delta_1$.

[11] The qualifier 'generically' is needed only because of the special case in which mimic costs happen to be exactly $\underline{C}$. In this case the proportion of cooperators in the stable steady state can be any value in the interval between zero and $p^{\text{EQ}}_{|C^{\text{Mim}}=\underline{C}}$. Regardless, the values of p^* $(= p^{\text{EQ}}_{|C^{\text{Mim}}=\underline{C}} = p^{\text{IM}}_{|C^{\text{Mim}}=\underline{C}})$ and v^* $(= u(0,0))$ remain unique in this special case.

Figure 2: Long Run Outcomes Within Consciousness Type



Notes. These figures illustrate how the long run conditional distribution (Panel A), π^*, the long run matching success (Panel B), p^*, and the economic payoff (Panel C), v^*, vary with mimic costs.

3. $1 \leq p^{\text{EP}}$. This is equivalent to $C^{\text{Mim}} \geq \overline{C}$. From (42), we have $\dot{\pi}(\pi) > 0$ for all π so that $\pi \to \pi^* = 1$ and $p \to p^* = 1$. The value of v^* is $u(1,1)$ since this is the payoff of cooperators at $p^* = 1$ (which is greater than the payoff of non-cooperator mimickers at $p^* = 1$).

Finally, $p^{\text{IM}} \geq 1$ implies $p^{\text{EP}} \geq p^{\text{IM}}$ under decreasing differences. Thus case 3 applies also when $p^{\text{IM}} \geq 1$.

Therefore, π^* is generically unique (taking the values derived above) and (π^*, p^*, v^*) takes the values described in the Lemma.

For completeness, the remaining (non-generic) case is $p^{\text{EP}} = p^{\text{IM}}$. This is equivalent to $C^{\text{Mim}} = \underline{C}$. From (42), we have $\dot{\pi}(\pi) < 0$ for $\pi \in (p^{\text{IM}}, 1]$ and $\dot{\pi}(\pi) = 0$ for $\pi \in [0, p^{\text{IM}}]$ so that $\pi \to \{[0, p^{\text{IM}}]\}$. The value of π^* is therefore not unique in this special case. Yet, the values of (p^*, v^*) remain unique: $p \to p^{\text{IM}} = p^{\text{EP}}$ and $v^* = u(0,0)$ since p^* leaves non-cooperators indifferent to mimicking. □

These outcomes are illustrated in Figure 2.

A.3 Evolution across q

Having derived long run values of (π, p, v) as a function of mimic costs in Lemma 6, we can now express these explicitly as functions of q. That is:

$$\pi_q^* \equiv \pi^*\left(C^{\text{Mim}}(q)\right), p_q^* \equiv p^*\left(C^{\text{Mim}}(q)\right), v_q^* \equiv v^*\left(C^{\text{Mim}}(q)\right). \tag{43}$$

From here it is straightforward to derive the long run fitness of q types:

$$f_q^* = v^*\left(C^{\text{Mim}}(q)\right) - C^{\text{Bio}}(q). \tag{44}$$

Deriving the dynamics of the marginal distribution of consciousness types then requires deriving the smallest local maximizer of f^*.

Proof of Proposition 1. Lemma 5 shows that strictly increasing differences or equal differences (generically) ensure $\pi^*(q) \in \{0,1\}$ for all q, including $q = 1$. The coexistence principle (Lemma 1) then implies $q^* \leq 1$, as required. □

Proof of Proposition 2. From Lemma 6, if $\kappa_0 \equiv C^{\text{Mim}}(1) \geq \overline{C}$ then $\pi_1^* = 1$. The coexistence principle (Lemma 1) then implies $q^* \leq 1$. Thus, a necessary condition for non-trivial consciousness is $\kappa_0 < \overline{C}$.

Another necessary condition for non-trivial consciousness is that minimally conscious agents are fitter than non-conscious agents. That is, we seek the conditions under which $f_0^* < f_1^*$. Since $f_0^* = u(0,0)$ and $f_1^* = v_1^* - \psi_0$, we seek the conditions under which:

$$v_1^* > u(0,0) + \psi_0. \tag{45}$$

From Lemma 6, if $\kappa_0 \leq \underline{C}$ then $v_1^* = u(0,0)$, which implies (45) can't hold in this case. Similarly, if $\kappa_0 \geq \overline{C}$ then $v_1^* = u(1,1)$. Thus, (45) will always hold in this case by the assumption that $\psi_0 < u(1,1) - u(0,0)$. Note that if this assumption were violated, then (45) will never hold and thus consciousness would never emerge.

If $C^{\text{Mim}}(1) \in \left(\underline{C}, \overline{C}\right)$ then Lemma 6 shows that $v_1^* = u(1,0) + p^{\text{EQ}} \cdot \delta_1$. Therefore, in this case, (45) holds if and only if

$$u(1,0) + \frac{\kappa_0 - (u(0,0) - u(1,0))}{\delta_0 - \delta_1} \cdot \delta_1 > u(0,0) + \psi_0. \tag{46}$$

Simplifying this gives $\kappa_0 > \underline{C}'$. Thus, another necessary condition for non-trivial consciousness is $\kappa_0 > \underline{C}'$. □

Proof of Proposition 4.

If $C^{\text{Mim}}(q) \leq \underline{C}$, then cooperators die out and $v_q^* = u(0,0)$ so that $v_q^* - v_0^* = 0$. Since $C^{\text{Mim}}(q) \leq \underline{C}$ implies $\Phi(q) = 0$, the result is proved for $C^{\text{Mim}}(q) \leq \underline{C}$.

Now consider $C^{\text{Mim}}(q) > \underline{C}$. In such cases there will exist some cooperators in the long run, implying that v_q^* will coincide with their economic payoff. That is:

$$v_q^* - v_1^* = u(1,0) + p_q^* \cdot \delta_1 - u(0,0) = p_q^* \cdot \delta_1 - \alpha_0, \tag{47}$$

where $\alpha_0 \equiv u(0,0) - u(1,0)$, and p_q^* is the long run match quality among q-consciousness types, which is given by:

$$p_q^* = \min\left\{p^{\text{EQ}}(q), 1\right\} \tag{48}$$

$$= \min\left\{\frac{C^{\text{Mim}}(q) - \alpha_0}{\delta_0 - \delta_1}, 1\right\} \tag{49}$$

$$= \frac{1}{\delta_0 - \delta_1} \cdot \min\left\{C^{\text{Mim}}(q) - \alpha_0, \delta_0 - \delta_1\right\} \tag{50}$$

$$= \frac{1}{\delta_0 - \delta_1} \cdot \min\left\{C^{\text{Mim}}(q), \delta_0 - \delta_1 + \alpha_0\right\} - \frac{\alpha_0}{\delta_0 - \delta_1} \tag{51}$$

$$= \frac{1}{\delta_0 - \delta_1} \cdot \min\left\{C^{\text{Mim}}(q), \overline{C}\right\} - \frac{\alpha_0}{\delta_0 - \delta_1}. \tag{52}$$

Therefore

$$v_q^* - v_1^* = \frac{\delta_1}{\delta_0 - \delta_1} \cdot \min\left\{C^{\text{Mim}}(q), \overline{C}\right\} - \frac{\delta_1}{\delta_0 - \delta_1} \cdot \frac{\delta_0}{\delta_1} \cdot \alpha_0 \tag{53}$$

$$= \frac{\delta_1}{\delta_0 - \delta_1} \cdot \left[\min\{C^{\text{Mim}}(q), \overline{C}\} - \underline{C}\right], \tag{54}$$

where the final line uses $\underline{C} \equiv \alpha_0 \cdot \frac{\delta_0}{\delta_1}$. This can be written as:

$$v_q^* - v_1^* = \frac{\delta_1}{\delta_0 - \delta_1} \cdot \left[\min\{C^{\text{Mim}}(q) - \underline{C}, \overline{C} - \underline{C}\}\right] \tag{55}$$

$$= \min\left\{\frac{\delta_1}{\delta_0 - \delta_1} \cdot (C^{\text{Mim}}(q) - \underline{C}), \frac{\delta_1}{\delta_0 - \delta_1} \cdot (\overline{C} - \underline{C})\right\} \tag{56}$$

$$= \min\{\phi(q), \Delta\}. \tag{57}$$

Therefore:

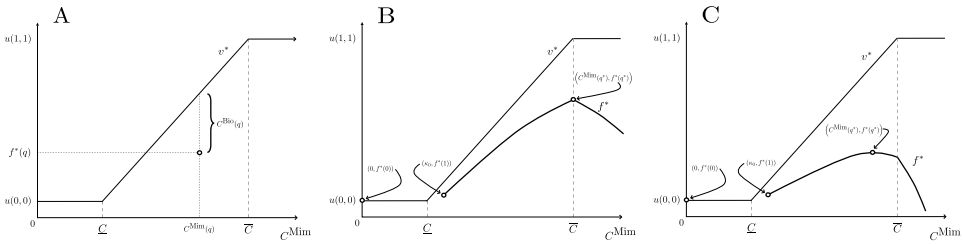
$$v_q^* - v_1^* = \begin{cases} 0 & \text{if } C^{\text{Mim}}(q) \leq \underline{C} \\ \min\{\phi(q), \Delta\} & \text{if } C^{\text{Mim}}(q) > \underline{C} \end{cases} \tag{58}$$

Since $C^{\text{Mim}}(q) \leq \underline{C}$ if and only if $\phi(q) \leq 0$, and since $\min\{\phi(q), \Delta\} \leq 0$ if and only if $\phi(q) \leq 0$ (because $\Delta > 0$), we can write:

$$v_q^* - v_1^* = \begin{cases} 0 & \text{if } \min\{\phi(q), \Delta\} \leq 0 \\ \min\{\phi(q), \Delta\} & \text{if } \min\{\phi(q), \Delta\} > 0, \end{cases} \tag{59}$$

which is equivalent to $\max\{0, \min\{\phi(q), \Delta\}\}$. □

Figure 3: Long Run Outcomes Across Consciousness Types



Notes. Panel A shows how $f^*(q)$ is derived given the mimic and biological costs associated with q. Panel B shows f^* as a function of mimic costs in the case where $\kappa_0 > \underline{C}'$ (i.e. $f^*(0) < f^*(1)$) and marginal biological costs, ψ, are low. Panel C shows f^* as a function of mimic costs in the case where $\kappa_0 > \underline{C}'$ (i.e. $f^*(0) < f^*(1)$) and marginal biological costs, ψ, are high.

The text describes a fundamental upper bound, $\bar{q}$, and Figure 3 illustrates how it constrains consciousness. The long run fitness of q types, given by (44), is depicted in Panel A. Panel B adds f_q^* as a function of q, and considers the case of low marginal biological costs. Here the smallest local maximizer of f_q^* is constrained at the upper bound, $\bar{q}$. Consciousness ceases evolving *not* because of biological costs to consciousness, but because there are no more fitness gains from superior matching at higher consciousness levels. Panel C, in contrast, traces the relationship between fitness and q when marginal biological costs are high. In this case the smallest local maximizer of f_q^* lies below the upper bound. Consciousness ceases evolving even though there continue to be payoff gains from superior matching. These gains simply come at too high a biological cost.

Proposition 5 provides a characterisation of long run consciousness.

Proposition 5 (Characterisation of Consciousness Level)**.** *Assume decreasing differences and $\kappa_0 > \underline{C}'$. The evolved level of consciousness, q^*, is positive and equal to the smallest local maximizer of:*

$$F(q) \equiv \frac{\delta_1}{\delta_0 - \delta_1} \cdot \min\{C^{Mim}(q), \overline{C}\} - C^{Bio}(q) \tag{60}$$

on $\mathbb{N}_1$. The evolved proportion of cooperators, $\pi^* \equiv \pi(q^*)$, is positive.

Proof. Since $C^{\text{Mim}}(1) > \underline{C}'$ and $\underline{C}' > \underline{C}$, we have $C^{\text{Mim}}(1) > \underline{C}$ and therefore $C^{\text{Mim}}(q) > \underline{C}$ for all $q \geq 1$. From Lemma 6), we therefore have

$$p^*(q) = \min\{p^{\text{EQ}}(C^{\text{Mim}}(q)), 1\} \tag{61}$$

since p^{EQ} is increasing in C^{Mim} with $p^{\mathrm{EQ}}(\overline{C}) = 1$. Therefore the economic payoff is

$$v^*(q) = u(1,0) + \min\{p^{\mathrm{EQ}}(C^{\mathrm{Mim}}(q)), 1\} \cdot \delta_1 \tag{62}$$

$$= u(1,0) + \min\{\frac{C^{\mathrm{Mim}}(q) - (u(0,0) - u(1,0))}{\delta_0 - \delta_1}, 1\} \cdot \delta_1 \tag{63}$$

$$= u(1,0) + \min\{C^{\mathrm{Mim}}(q) - (u(0,0) - u(1,0)), \delta_0 - \delta_1\} \cdot \frac{\delta_1}{\delta_0 - \delta_1} \tag{64}$$

$$= u(1,0) - \frac{u(0,0) - u(1,0)}{\delta_0 - \delta_1} + \min\{C^{\mathrm{Mim}}(q), \delta_0 - \delta_1 + u(0,0) - u(1,0)\} \cdot \frac{\delta_1}{\delta_0 - \delta_1} \tag{65}$$

$$= u(1,0) - \frac{u(0,0) - u(1,0)}{\delta_0 - \delta_1} + \min\{C^{\mathrm{Mim}}(q), u(0,1) - u(1,1)\} \cdot \frac{\delta_1}{\delta_0 - \delta_1} \tag{66}$$

$$= \bar{V} + V(q), \tag{67}$$

where $\bar{V} \equiv u(1,0) - \frac{u(0,0)-u(1,0)}{\delta_0-\delta_1}$ is independent of q and $V(q) \equiv \min\{C^{\mathrm{Mim}}(q), \overline{C}\} \cdot \frac{\delta_1}{\delta_0-\delta_1}$. Therefore fitness can be written

$$f^*(q) = v^*(q) - C^{\mathrm{Bio}}(q) = \bar{V} + F(q), \tag{68}$$

where $F(q)$ is defined in the proposition. Since $\bar{V}$ is independent of q, the smallest local maximizer of f^* is the smallest local maximizer of $F(q)$. The smallest local maximizer is positive by virtue of $C^{\mathrm{Mim}} > \underline{C}'$ and Proposition 2. The evolved proportion of cooperators, $\pi^* \equiv \pi^*_{q^*} = \min\{p^{\mathrm{EQ}}(C^{\mathrm{Mim}}(q^*)), 1\}$, is positive by virtue of $C^{\mathrm{Mim}} > \underline{C}$ and Lemma 6. □

Proof of Proposition 3. The claim holds trivially for $\tilde{q} = 0$, so consider some $\tilde{q} \geq 1$. The proof is by construction.

First, noting that the assumption $\psi_0 < u(1,1) - u(0,0)$ implies $\underline{C}' < \overline{C}$, choose (κ_0, ψ_0) such that $C^{\mathrm{Mim}}(1) \in \left(\underline{C}', \overline{C}\right)$. That is, such that:

$$\underline{C} + \frac{\delta_0 - \delta_1}{\delta_1} \cdot \psi_0 < \kappa_0 < \overline{C}. \tag{69}$$

Since $\underline{C} < \overline{C}$, this condition can always be satisfied (e.g. choose any $\kappa_0 \in (\underline{C}, \overline{C})$ and then choose ψ_0 sufficiently small). Note too that the condition ensures $f^*(1) > f^*(0)$ since $C^{\mathrm{Mim}}(1) > \underline{C}'$. Thus, this condition ensures the proposition holds for $\tilde{q} = 1$. From here, then, consider $\tilde{q} \geq 2$.

Second, choose κ so that $\pi(\tilde{q})$ is interior: $C^{\text{Mim}}(\tilde{q}) < \overline{C}$. For $\tilde{q} \geq 2$, we need to choose κ so that:

$$\kappa < \underline{\kappa}(\kappa_0, \tilde{q}) \equiv \frac{\overline{C} - \kappa_0}{c(\tilde{q}-1)}. \tag{70}$$

We have $\underline{\kappa}(\kappa_0, \tilde{q}) > 0$ since $\overline{C} > \kappa_0$ by (69), and therefore this condition can always be satisfied for positive κ by making it sufficiently small.

Third, choose ψ so that $\frac{d}{dq}F^*(q) > 0$ for all $q \in [1, \tilde{q}]$:

$$\frac{\delta_1}{\delta_0 - \delta_1} \cdot \kappa \cdot c'(q-1) > \psi \cdot b'(q-1)$$

for all $q \in [1, \tilde{q}]$. That is:

$$\psi < \overline{\psi}(\kappa, \tilde{q}) \equiv \kappa \cdot \frac{\delta_1}{\delta_0 - \delta_1} \cdot \Omega(\tilde{q}), \tag{71}$$

where

$$\Omega(\tilde{q}) \equiv \max_{q \in [1, \tilde{q}]} \frac{c'(q-1)}{b'(q-1)}.$$

Since $\overline{\psi}(\kappa, \tilde{q}) > 0$ for any $\kappa > 0$ and any $\tilde{q} \in [1, \infty)$, this can always be satisfied by positive ψ by making it sufficiently small.

In summary, choosing $(\{\kappa_0, \kappa\}, \{\psi_0, \psi\})$ so that (69), (70), and (71) are satisfied ensures $f^*(1) > f^*(0)$ and $f^*(\tilde{q}) > f^*(q)$ for all $q \in \{1, 2, ..., \tilde{q}\}$. The smallest local maximizer of $F^*(q)$ is therefore at least $\tilde{q}$. □

Lemma 7. *Assume decreasing differences and $\kappa_0 \geq \underline{C}'$. The evolved level of consciousness, q^* is:*

1. *weakly decreasing in the marginal biological cost parameter, ψ.*
2. *hump-shaped in the marginal mimic cost parameter, κ.*

Proof. We consider the effect of marginal biological and mimicking costs in turn as follows.

Marginal Biological Costs From Proposition 5, the value of q^* is the smallest local maximizer of $F(q) = V(q) - C^{\text{Bio}}(q)$, where

$$V(q) \equiv \frac{\delta_1}{\delta_0 - \delta_1} \cdot \min\{C^{\text{Mim}}(q), \overline{C}\}. \tag{72}$$

That is, q^* satisfies:

$$V(q^*) - V(q) > \psi \cdot [b(q^* - 1) - b(q - 1)] \ \forall q \in \{1, ..., q^* - 1\}, \text{ and} \tag{73}$$
$$V(q^* + 1) - V(q^*) \leq \psi \cdot [b(q^*) - b(q^* - 1)]. \tag{74}$$

A decrease in ψ will preserve inequality (73), thereby showing that q^* cannot decrease as a result. That is, q^* is non-increasing in ψ. Yet, a sufficient decrease in ψ will violate inequality (74) when $q^* < \bar{q}$ (since the left hand side is strictly positive in this case), thereby showing that q^* will eventually increase as ψ decreases in such cases.

As $\psi \to 0$, we have $q^* = \bar{q}$ since F becomes a strictly increasing function on $\{1, ..., \bar{q}\}$. As $\psi \to \infty$ we have $F(1) \geq F(q)$ for all $q \geq 1$ and therefore the smallest local maximizer goes to $q = 1$.

Marginal Mimic Costs From Proposition 5, the value of q^* is the smallest local maximizer of

$$F(q|\kappa) \equiv \frac{\delta_1}{\delta_0 - \delta_1} \cdot \min\{C^{\text{Mim}}(q|\kappa), \overline{C}\} - C^{\text{Bio}}(q) \tag{75}$$

on $\mathbb{N}_1$. Let $\bar{q}(\kappa) \equiv \min\{q |\ C^{\text{Mim}}(q|\kappa) \geq \overline{C}\}$, noting that it is decreasing in κ. In particular, $\kappa_0 \geq \overline{C}$ implies $\bar{q}(\kappa) = 1$ for all κ. If $\kappa_0 < \overline{C}$ then $\bar{q}(\kappa) = \tilde{q}$ if $\kappa \in \left[\kappa^{[\tilde{q}]}, \kappa^{[\tilde{q}-1]}\right)$ where $\kappa^{[1]} \equiv \infty$ and $\kappa^{[q]} \equiv [\overline{C} - \kappa_0]/c(q-1)$ for $q \geq 2$.

Let $q^{**}(\kappa)$ be the smallest local maximizer of

$$\mathcal{F}(q) \equiv \frac{\delta_1}{\delta_0 - \delta_1} \cdot C^{\text{Mim}}(q|\kappa) - C^{\text{Bio}}(q) \tag{76}$$

on $\mathbb{N}_1$, noting that it is increasing in κ with $q^{**}(0) = 1$.

If κ is low enough that $q^{**}(\kappa) < \bar{q}(\kappa)$, then clearly $q^*(\kappa) = q^{**}(\kappa)$ (if the unconstrained optimal is feasible it is also constrained optimal). Thus $q^*(\kappa)$ is increasing in this region.

For higher values of κ, i.e. such that $q^{**}(\kappa) \geq \bar{q}(\kappa)$, $q^*(\kappa)$ is either $\bar{q}(\kappa)$ or $\bar{q}(\kappa) - 1$ depending on which has the higher fitness ($q^*(\kappa)$ cannot be larger than $\bar{q}(\kappa)$ since the latter is an upper bound, and $q^*(\kappa)$ cannot be smaller

than $\bar{q}(\kappa) - 1$ since $q^{**}(\kappa) > \bar{q}(\kappa) - 1)$. Specifically, $\bar{q}(\kappa) - 1$ has the higher fitness if:

$$\frac{\delta_1}{\delta_0 - \delta_1} \cdot [\overline{C} - C^{\text{Mim}}(q|\kappa)] < C^{\text{Bio}}(\bar{q}(\kappa)) - C^{\text{Bio}}(\bar{q}(\kappa) - 1). \tag{77}$$

If this holds at some $\kappa \in \left[\kappa^{[\bar{q}]}, \kappa^{[\bar{q}-1]}\right)$, then it will also hold for all higher values in that interval, implying $q^*(\kappa)$ is weakly decreasing on this interval. If κ increases further, so that $\bar{q}(\kappa)$ falls, then $q^*(\kappa)$ must be non-decreasing (in the original interval $q^*(\kappa) \geq \bar{q}(\kappa) - 1$ and in the next interval $q^*(\kappa') \leq \bar{q}(\kappa') = \bar{q}(\kappa) - 1$, implying $q^*(\kappa') \leq q^*(\kappa)$). Either way, $q^*(\kappa)$ is weakly decreasing in this region. That is, $q^*(\kappa)$ is hump-shaped.

To summarize, if $\kappa_0 \geq \overline{C}$ then $q^*(\kappa) = 1$ for all κ. If $\kappa_0 < \overline{C}$ let $\hat{\kappa} \equiv \max\{\kappa \mid q^{**}(\kappa) < \bar{q}(\kappa)\}$. Then $q^*(\kappa)$ is weakly increasing for $\kappa < \hat{\kappa}$, is weakly decreasing for $\kappa > \hat{\kappa}$, with $\lim_{\kappa \downarrow \hat{\kappa}} q^*(\kappa) \in \{q^*(\hat{\kappa}), q^*(\hat{\kappa}) + 1\}$. □

B Discussion of Literature on Role of Consciousness

The claim that there is not yet a convincing account of the role of consciousness collides, at least initially, with strong intuitions to the contrary. Here we briefly outline existing arguments made by scientists and philosophers supporting the consensus view that those intuitions are implausible.

Most of us have an introspective notion that consciousness seems to play an important causal role in our behavior. For instance, we have the sense that consciousness provides the impetus for appropriate responses; e.g. feeling pain is needed so that we know to pull our hand out of the fire, feeling pleasure is needed so that we pursue calories and sex, feeling fear is needed so that we know to flee a hungry lion, and so on. It certainly *seems* that this is the role of consciousness, even if we scoff at the idea of a homunculus sitting in the Cartesian Theatre (see Dennett (1991)). Yet there are various ways to see that this class of explanation is highly implausible.

On logical grounds, "produce experience" cannot usefully be part of a causal chain connecting stimulus and response. Since experience is generated within the brain, any inner state that is experienced is merely a pattern of neuronal firing triggered by some earlier neuronal firing that can eventually be traced back to an initiating stimulus. Any behavior following downstream from experience is only, at best, proximately caused by experience. As Gutfreund (2018) puts it: "If behavior is caused fully by unconscious

neural circuits, how can it also be caused by feelings?"[12] Whatever beneficial behaviors are claimed to follow from inner experiences could have equivalently, and more directly, followed from the neuronal activation preceding the inner experience. Since the ultimate benefit for the organism is behavior, this could have been gained most directly by simply bypassing any excursion through inner experiences on the way to behavior.

On empirical grounds, there is no place in the brain where input signals are collected and processed into behavioural responses (no homunculus), and there is much evidence suggesting 'consciousness is the last to know'; i.e., experience arises *after* the relevant neuronal response has been initiated. Many studies have confirmed Libet (1978) surprising findings suggesting that action activation precedes consciousness; see Soon et al. (2008) and Soon et al. (2013) for FMRI evidence.

An alternative explanation for the value of consciousness may be because it *necessarily* accompanies behaviors or processes that enhance fitness. Processes such as integrating input data from multiple senses, directing attention, forming a theory of mind, engaging in deliberative thinking, metacognition, and so on, are clearly of survival value. Most of the leading 'theories of consciousness' take one of these, or another (see Seth and Bayne (2022) for a survey), and treat first-person experience as a necessary by-product of the chosen cognitive function. Here the issue is explaining why the experiential component *per se* is playing any role; a point made by both Block (1995) and Chalmers (1996).

To illustrate, consider one of the most popular classes of theories – Global Workspace Theories (GWTs). Its core claim is that when sensory information is moved from pre- or sub-conscious systems to conscious ones it essentially moves into a "global workspace" within the brain. In the workspace information becomes accessible to multiple operating modules that can utilize it to guide behavior. Support for aspects of the theory comes from the lab. When participants see objects without conscious realization of seeing, neuronal firing is limited to areas of the visual cortex. However, when they *experience* "seeing the object", neuronal activation extends beyond the visual cortex, synchronizing firing there with the firing of multiple additional cortical regions. So the occurrence of experience coincides with localized information in the brain seemingly becoming accessible to processes in the brain far beyond.

Clearly if having an experience, i.e., being conscious, makes information more widely available in the brain it could provide a fitness advan-

[12] See also Harari (2016) for an intuitive elaboration of this argument.

tage. But even if GWT is correct, and making information broadly available in the brain coincides with consciousness, this theory does not explain why experience, or being conscious, needs to accompany the making of information broadly available. The problem with such explanations, as Chalmers (1995) notes, is that consciousness – having subjective experiences – does not seem necessary for any of them. Theories like this do not attempt to explain the function of first-person experience. Rather, they attach experience to some function performed by the brain and then detail the function. This yields no insight into the evolutionary benefit of experience, even if an evolutionary benefit of the function is clear. [13]

It is immediately clear that attaching experience to complex processing in this way must be ad hoc because there are many examples of systems with immense functionality but without any first-person experiential accompaniment. Humans have built such systems, e.g. self-driving cars, without the need to program in experience. Moreover, this is the case for most human functions, for example: the regulation of salt levels, of hormones, of digestion, of temperature, of heart rate, which are all functions performed without consciousness.[14] As Velmans (2014) notes:

> "Cognitive theories which identify consciousness with one or another information processing "box" simply assume or define it to be ontologically identical to a given form of processing in the brain (largely ignoring its phenomenology). Such theories typically move, without blinking, from relatively well-justified claims about the forms of information processing with which consciousness is associated, to entirely unjustified claims about what consciousness is or what it does...., such manoeuvres beg the question; that is, they assume or posit what they need to establish."

For a theory to explain why consciousness, as subjective experience, evolved, either it must explain the fitness enhancing contribution of the experiencing part per se, or it must explain why experience *necessarily* accompanies a fitness enhancing function. As yet, no theory does either.

[13] Other examples subject to the same criticism are: Integrated Information Theory that proposes consciousness arises when information processing takes a particular (integrated) form (Tononi et al. (2016)), Unlimited Associative Learning (Birch et al. (2020)), which posits a type of learning uniquely accompanying conscious beings, Complex Decision Making (Earl (2014)), where complexity for some reason requires consciousness.

[14] See Chalmers (2004) for further elaboration. The share of cognitive functions occurring without experience is well in excess of those that are conscious; see Bargh and Morsella (2008).

C Evidence in support of model assumptions

We present some evidence drawn from psychological literature in support of key building blocks in the analysis.

C.1 Organisms differ by type

Assumption: *Organisms within a species may differ by "type". Types are not directly observable, and predict actions that are payoff relevant for other organisms.*

Psychologists have categorized a set of personality traits claimed to predict much of the variation in cross individual behavior. These are denoted the "Big Five" personality traits: extraversion, agreeableness, openness, conscientiousness, and neuroticism. It is claimed individual propensities can be detected as early as age three, and become more stable as children develop into adults (Caspi et al. (2005). Though population means can change with age, there tends to be high rank-order consistency within cohorts through time; see Roberts et al. (2006).

With respect to cooperative behavior specifically, Peysakhovich et al. (2014) report a large degree of cross domain, within individual, through time stability in decisions regarding cooperative behavior. They dub this the "cooperative phenotype". Heterogeneity in cooperative behavior has been well documented in strategic games; see, for example, Fischbacher and Gächter (2010). And the interpretation that behavioural differences in observed cooperation reflect a stable dispositional type is reinforced by FMRI studies that have detected type specific neurosignatures; see Gianotti et al. (2019).[15]

In our formal model we simplify to humans coming in two types – cooperators and non-cooperators. In the neuropsychology literature, some studies have observed honesty as a default requiring cognitive effort to override, and others the opposite, honesty requiring increases in cognitive effort. A reconciliation proposed by Speer et al. (2020) suggests underlying type heterogeneity may be the explanation. They demonstrate that habitual liars have different forms of neural activation to habitually honest individuals. In an incentivized task where subjects could lie for reward, they

[15] Baumgartner et al. (2019) report that individuals sorted by degree of cooperativeness via incentivized interactions (unconditional cooperators, conditional cooperators and non-cooperators) have observable differences in neuronal baseline activation. They systematically differed in Temporoparietal junction (TPJ), and left Lateral Pre-Frontal Cortex activation even in resting state.

showed this to be the case. Individual brain areas associated with cognitive control (anterior cingulate cortex and inferior frontal gyrus) helped habitually dishonest participants to be honest, whereas these enabled cheating for usually honest participants.

A physiological predicter of individual level deceptiveness was reported by Baumgartner et al. (2013). Using EEG measurements they found task-independent baseline activation in the anterior insula, a brain area implicated in mapping internal bodily states and in representing emotional arousal and conscious feelings, predicts individuals' propensity for deceptive behavior. EEG signatures are stable over time – suggesting a good correspondence with the behavioral stability of a 'type' that we use in our model, as is typical in economics. The authors speculate as to the reasons for the underlying heterogeneity in subjects neural baseline activation that were predictive of the effects of cognitive control on lying behavior, some of which are undoubtedly genetic, but may also be learned, or a combination. In general, it seems that this heterogeneity in individuals also makes the use of positive or negative affect a poor predictor of lying activity, see Gamer and Suchotzki (2018) for a further discussion. The literature does not weigh in decisively on the relative role of biological or social constructions contributing to type, and it accordingly does not matter to our theory. Though it is simple to model genetic evolution of a behavioral type in our theory, it can be easily respecified in a cultural evolution, or an indirect evolutionary formalization.

The Big Five traits are predictive of payoff relevant personality types. For example, the trait 'Agreeableness' has been consistently found to positively associate with altruism.(Ashton and Lee (2008); Graziano et al. (1997). 'Openness' has been shown to correlate with pro-social behavior (Van Lange et al. (1997) as has 'Conscientiousness' (Graziano et al. (1997)). Studies have also shown that both 'Neuroticism' and 'Extroversion' correlate with altruism (negative for neuroticism; Bekkers and Wiepking (2011), and positive for extroversion).

C.2 Types vary by experience

Assumption: *The content of experienced inners states can, in principle, differ by type. So two organisms of different types subject to the same set of external conditions, will experience distinct internal states for some subset of external conditions. And, reciprocally, the internal states experienced by organisms of the same type are the same over some domains of experience.*

In the context of our two type example, this assumption asserts that the

cooperator type experiences a distinct inner state from the non-cooperator. For example, the cooperator experiences inner turmoil, guilt or pain when causing harm to others, a non-cooperator experiences no such inner turmoil, or perhaps a mitigated form of it.

In the Baumgartner et al. (2013) study discussed above, in addition to the EEG measurements of baseline activation in the anterior insula, subjects also reported their personal propensity to experience negative emotions. They found this phenomenological report to correlate with cheating and type. In short, they suggest that subjects with high baseline activation in the anterior insula might avoid the deceptive act because their emotional system would react too strongly in such an aversive situation.

Type, in the broader sense of individuals with different personality dispositions, also seems to incline individuals to distinct experiences. A number of Big 5 traits have been shown to correlate with distinct conscious experiences. Larsen et al. (1986) found that individuals scoring high on the neuroticism dimension reported greater intensity of negative emotional experiences under identical emotional stimulus. Extraversion was found to positively correlate with experiences of happiness, controlling for life situation (Lucas and Diener (2008)). Conscientiousness correlates with inner feelings of pride in subjects upon accomplishment of goals (Roberts et al. (2004)). Another Big Five factor, openness, was shown to correlate with distinct perceptual experiences such as synesthesia, (DeYoung et al. (2002)).

C.3 Messaging inexperienced content is costly

Assumption: *Messaging the contents of experienced inner states to other organisms is less costly/easier than messaging the contents of inner states that are not directly experienced.*

One reason for the cost of messaging inner states to be higher when they are not directly experienced is because the messenger must learn the contents of the other type's inner state to message it. Other reasons would arise if there are metabolic costs arising from the cognitive effort of asserting the experience of inner states that are false. There seem to be costs to inauthentic reporting of an individual's internal states including negative affect, psychological distress and physiological arousal; which is costly metabolicaly. Falsely reporting the contents of emotional states led to experiences of greater negative affect than when providing authentic reports; Feldman Barrett and Russell (1998). Vansteenkiste et al. (2004) find that false reporters experienced greater psychological distress and lower well-being; and Gross and Levenson (1997) show that false reporting about emo-

tional state also lead to greater physiological arousal and negative affect than expressing true emotions.

The inner experiences of altruists appear to differ from individuals who undertake altruistic appearing behavior but are not truly altruists. Strategic altruists, by observing social norms and expectations regarding prosocial behavior and intuiting the required behaviors in particular contexts, use these to guide their own behavior in order to develop altruistic reputations. Ochsner and Gross (2008) describe how such strategic agents attempt to curate their inner experiences away from default settings by engaging in 'cognitive reappraisal'. That is they actively change their own thoughts and attitudes about a situation in order to appear more altruistic. For example, they may consciously shift their focus away from their own self-interest and toward the well-being of others in order to make their actions appear more genuinely altruistic. They may also engage in emotional regulation strategies to appear more altruistic by suppressing negative emotions or amplifying positive ones to create a more convincing altruistic persona. These authors provide an overview of cognitive emotion regulation strategies and discuss the neural mechanisms that underlie these processes.

So inauthentic reports of one's own conscious state require an individual to first learn the conscious state of the individual type they are trying to mimic, which is in itself costly. Secondly acts of suppressing or reappraising ones own emotions, have been well documented to lead to negative affect, and considerable evidence of increased arousal; both of which are likely to involve metabolic costs.

References

Alger, I. (2023). Evolutionarily stable preferences [Publisher: Royal Society]. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *378*(1876), 20210505. https://doi.org/10.1098/rstb.2021.0505 (cit. on p. 5)

Alger, I., & Weibull, J. W. (2013). Homo Moralis—Preference Evolution Under Incomplete Information and Assortative Matching [Publisher: [Wiley, Econometric Society]]. *Econometrica*, *81*(6), 2269–2302. Retrieved November 30, 2023, from https://www.jstor.org/stable/23524319 (cit. on p. 5)

Ashton, M. C., & Lee, K. (2008). The prediction of Honesty-Humility-related criteria by the HEXACO and Five-Factor Models of personality [Place:

Netherlands Publisher: Elsevier Science]. *Journal of Research in Personality*, *42*(5), 1216–1228. https://doi.org/10.1016/j.jrp.2008.03.006 (cit. on p. 44)

Bargh, J., & Morsella, E. (2008). The unconscious mind. *Perspectives on Psychological Science*, *3(1)* (cit. on p. 42).

Barlow, H. (1997). Single neurons, communal goals, and consciousness. In M. Ito, Y. Miyashita, & E. Rolls (Eds.), *Cognition, computation, and consciousness* (pp. 121–136). Oxford University Press. (Cit. on p. 3).

Baumgartner, T., Gianotti, L. R. R., & Knoch, D. (2013). Who is honest and why: Baseline activation in anterior insula predicts inter-individual differences in deceptive behavior. *Biological Psychology*, *94(1)*, 192–197 (cit. on pp. 44, 45).

Baumgartner, T., Dahinden, F. M., Gianotti, L. R. R., & Knoch, D. (2019). Neural traits characterize unconditional cooperators, conditional cooperators, and noncooperators in group-based cooperation. *Human Brain Mapping*, *40*(15), 4508–4517. https://doi.org/10.1002/hbm.24717 (cit. on p. 43)

Bekkers, R., & Wiepking, P. (2011). A Literature Review of Empirical Studies of Philanthropy: Eight Mechanisms That Drive Charitable Giving. *Nonprofit and Voluntary Sector Quarterly*, *40*(5), 924–973. https://doi.org/10.1177/0899764010380927 (cit. on p. 44)

Bidner, C. (2010). Pre-match investment with frictions. *Games and Economic Behavior*, *68*(1), 23–34. http://ideas.repec.org/a/eee/gamebe/v68y2010i1p23-34.html (cit. on p. 4)

Bidner, C. (2014). A spillover-based theory of credentialism. *Canadian Journal of Economics*, *47*(4) (cit. on p. 4).

Birch, J., Ginsburg, S., & Jablonka, E. (2020). Unlimited Associative Learning and the origins of consciousness: A primer and some predictions. *Biology and Philosophy*, *35* (cit. on p. 42).

Blackmore, S., & Troscianko, E. T. (2018). *Consciousness: An Introduction*. Taylor & Francis Group. Retrieved May 30, 2024, from http://ebookcentral.proquest.com/lib/sfu-ebooks/detail.action?docID=5352149. (Cit. on p. 2)

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, *18*(2), 227–247. https://doi.org/10.1017/S0140525X00038188 (cit. on pp. 1, 41)

Caspi, A., Roberts, B. W., & Shiner, R. L. (2005). Personality development: Stability and change. *Annual Review of Psychology*, *56*, 453–484. https://doi.org/10.1146/annurev.psych.55.090902.141913 (cit. on p. 43)

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2*(3), 200–219 (cit. on pp. 2, 19, 42).

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, Incorporated. Retrieved December 5, 2023, from http://ebookcentral.proquest.com/lib/sfu-ebooks/detail.action?docID=272854. (Cit. on pp. 6, 41)

Chalmers, D. J. (2004). How Can We Construct a Science of Consciousness? In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences Iii* (pp. 1111–1119). MIT Press. (Cit. on p. 42).

Chalmers, D. J. (2010). *The Character of Consciousness*. Oxford University Press. (Cit. on p. 19).

Churchland, P. M. (2013). *Matter and consciousness: A contemporary introduction to the philosophy of mind* (3rd ed.). MIT Press. https://mitpress.mit.edu/9780262519588/matter-and-consciousness/. (Cit. on p. 2)

Cole, H. L., Mailath, G. J., & Postlewaite, A. (1995). Incorporating concern for relative wealth into economic models. *Quarterly Review*, *19*(Sum), 12–21. https://ideas.repec.org/a/fip/fedmqr/y1995isump12-21nv.19no.3.html (cit. on p. 4)

Dehaene, S. (2014). *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. Viking. (Cit. on pp. 21, 25).

Dekel, E., Ely, J. C., & Yilankaya, O. (2007). Evolution of Preferences [Publisher: [Oxford University Press, Review of Economic Studies, Ltd.]]. *The Review of Economic Studies*, *74*(3), 685–704. Retrieved December 5, 2023, from https://www.jstor.org/stable/4626157 (cit. on p. 5)

Dennett, D. (1991). *Consciousness explained*. Little Brown; Co. (Cit. on pp. 2, 40).

DeYoung, C. G., Peterson, J. B., & Higgins, D. M. (2002). Higher-order factors of the Big Five predict conformity: Are there neuroses of health? [Place: Netherlands Publisher: Elsevier Science]. *Personality and Individual Differences*, *33*(4), 533–552. https://doi.org/10.1016/S0191-8869(01)00171-4 (cit. on p. 45)

Earl, B. (2014). The biological function of consciousness. *Frontiers in Psychology*, *5* (cit. on p. 42).

Feldman Barrett, L., & Russell, J. A. (1998). Independence and bipolarity in the structure of current affect [Place: US Publisher: American Psychological Association]. *Journal of Personality and Social Psychology*, *74*(4), 967–984. https://doi.org/10.1037/0022-3514.74.4.967 (cit. on p. 45)

Fischbacher, U., & Gächter, S. (2010). Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments [Publisher: American Economic Association]. *The American Economic Review*, *100*(1), 541–556. Retrieved December 5, 2023, from https://www.jstor.org/stable/27804940 (cit. on p. 43)

Fodor, J. (2004). You can't argue with a novel [ISBN: 9780262122597 Section: Literature & Criticism reviewed-title: Radiant Cool: A Novel Theory of Consciousness]. *London Review of Books*, *26*(05). Retrieved June 6, 2024, from https://www.lrb.co.uk/the-paper/v26/n05/jerry-fodor/you-can-t-argue-with-a-novel (cit. on p. 2)

Frank, R. H. (1987). If Homo Economicus Could Choose His Own Utility Function, Would He Want One with a Conscience? [Publisher: American Economic Association]. *The American Economic Review*, *77*(4), 593–604. Retrieved November 30, 2023, from https://www.jstor.org/stable/1814533 (cit. on pp. 5, 20)

Gamer, M., & Suchotzki, K. (2018). Lying and psychology. In J. Maibauer (Ed.), *The "oxford handbook of lying*. MIT Press. (Cit. on p. 44).

Gianotti, L. R. R., Dahinden, F. M., Baumgartner, T., & Knoch, D. (2019). Understanding Individual Differences in Domain-General Prosociality: A Resting EEG Study. *Brain Topography*, *32*(1), 118–126. https://doi.org/10.1007/s10548-018-0679-y (cit. on p. 43)

Graziano, W. G., Jensen-Campbell, L. A., & Finch, J. F. (1997). The self as a mediator between personality and adjustment [Place: US Publisher: American Psychological Association]. *Journal of Personality and Social Psychology*, *73*(2), 392–404. https://doi.org/10.1037/0022-3514.73.2.392 (cit. on p. 44)

Gross, J. J., & Levenson, R. W. (1997). Hiding feelings: The acute effects of inhibiting negative and positive emotion [Place: US Publisher: American Psychological Association]. *Journal of Abnormal Psychology*, *106*(1), 95–103. https://doi.org/10.1037/0021-843X.106.1.95 (cit. on p. 45)

Gutfreund, Y. (2018). The mind-evolution problem: The difficulty of fitting consciousness in an evolutionary framework. *Frontiers in Psychology* (cit. on p. 40).

Harari, Y. N. (2016). *Homo Deus: A brief history of tomorrow*. Harvill Secker. (Cit. on p. 41).

Heller, Y., & Mohlin, E. (2019). Coevolution of deception and preferences: Darwin and Nash meet Machiavelli. *Games and Economic Behavior*, *113*, 223–247. https://doi.org/10.1016/j.geb.2018.09.011 (cit. on p. 5)

Hopkins, E. (2012). Job market signaling of relative position, or becker married to spence. *Journal of the European Economic Association*, *10*(2), 290–322. https://doi.org/https://doi.org/10.1111/j.1542-4774.2010.01047.x (cit. on p. 4)

Hopkins, E. (2014). Competitive Altruism, Mentalizing, and Signaling. *American Economic Journal: Microeconomics*, *6*(4), 272–292. https://doi.org/10.1257/mic.6.4.272 (cit. on p. 5)

Hoppe, H. C., Moldovanu, B., & Sela, A. (2009). The theory of assortative matching based on costly signals. *The Review of Economic Studies*, *76*(1), 253–281. https://doi.org/10.1111/j.1467-937X.2008.00517.x (cit. on p. 4)

Humphrey, N. (2023). *Sentience: The invention of consciousness*. MIT Press. (Cit. on p. 3).

Kartik, N. (2009). Strategic Communication with Lying Costs. *The Review of Economic Studies*, *76*(4), 1359–1395. https://doi.org/10.1111/j.1467-937X.2009.00559.x (cit. on pp. 5, 20)

Kay, T., Keller, L., & Lehmann, L. (2020). The evolution of altruism and the serial rediscovery of the role of relatedness. *Proceedings of the National Academy of Sciences* (cit. on p. 5).

Kleiner, J. (2020). Mathematical Models of Consciousness. *Entropy (Basel)*, *30(6)* (cit. on p. 22).

Larsen, R. J., Diener, E., & Emmons, R. A. (1986). Affect intensity and reactions to daily life events [Place: US Publisher: American Psychological Association]. *Journal of Personality and Social Psychology*, *51*(4), 803–814. https://doi.org/10.1037/0022-3514.51.4.803 (cit. on p. 45)

Léon, D. d. (1997). *The Qualities of Qualia*. (Cit. on p. 20).

Libet, B. W. (1978). Neuronal Vs. Subjective Timing for a Conscious Sensory Experience. In P. A. Buser & A. Rougeul-Buser (Eds.), *Cerebral Correlates of Conscious Experience*. Elsevier. (Cit. on p. 41).

Lucas, R. E., & Diener, E. (2008). Personality and subjective well-being. In *Handbook of personality: Theory and research, 3rd ed* (pp. 795–814). The Guilford Press. (Cit. on p. 45).

Nagel, T. (1974). What Is It Like to Be a Bat? [Publisher: [Duke University Press, Philosophical Review]]. *The Philosophical Review*, *83*(4), 435–450. https://doi.org/10.2307/2183914 (cit. on pp. 1, 5)

Nagel, T. (2012). *Mind and cosmos: Why the materialist neo-darwinian conception of nature is almost certainly false*. Oxford University Press. (Cit. on p. 2).

Ochsner, K. N., & Gross, J. J. (2008). Cognitive Emotion Regulation: Insights from Social Cognitive and Affective Neuroscience. *Current directions in psychological science*, *17*(2), 153–158. https://doi.org/10.1111/j.1467-8721.2008.00566.x (cit. on p. 46)

Peysakhovich, A., Nowak, M. A., & Rand, D. G. (2014). Humans display a 'cooperative phenotype' that is domain general and temporally stable. *Nature Communications*, *5*(1), 4939. https://doi.org/10.1038/ncomms5939 (cit. on p. 43)

Roberts, B. W., O'Donnell, M., & Robins, R. W. (2004). Goal and Personality Trait Development in Emerging Adulthood [Place: US Publisher: American Psychological Association]. *Journal of Personality and Social Psychology*, *87*(4), 541–550. https://doi.org/10.1037/0022-3514.87.4.541 (cit. on p. 45)

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin*, *132*(1), 1–25. https://doi.org/10.1037/0033-2909.132.1.1 (cit. on p. 43)

Robson, A. J. (1990). Efficiency in evolutionary games: Darwin, nash and the secret handshake. *Journal of Theoretical Biology*, *144*(3), 379–396. https://doi.org/https://doi.org/10.1016/S0022-5193(05)80082-7 (cit. on p. 5)

Robson, A. J., & Samuelson, L. (2011). The evolutionary foundations of preferences. In J. Benhabib, A. Bisin, & M. O. Jackson (Eds.), *Handbook of Social Economics* (pp. 221–310). North-Holland. https://doi.org/10.1016/B978-0-444-53187-2.00007-3. (Cit. on p. 5)

Seth, A., & Bayne, T. (2022). Theories of consciousness. *Nature Reviews: Neuroscience* (cit. on p. 41).

Soon, C., Brass, M., Heinze, H.-J., & Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain. *Nature neuroscience*, *11*, 543–5. https://doi.org/10.1038/nn.2112 (cit. on p. 41)

Soon, C., Namburi, P., & Chee, M. W. L. (2013). Preparatory patterns of neural activity predict visual category search speed. *NeuroImage*, *66*, 215–222. https://doi.org/https://doi.org/10.1016/j.neuroimage.2012.10.036 (cit. on p. 41)

Speer, S., Smidts, A., & Boksem, M. (2020). Cognitive control increases honesty in cheaters but cheating in those who are honest. *Proceedings of the National Academy of Science*, *117 (32)*, 19080–19091 (cit. on p. 43).

Tononi, G., Boly, M., & Massimini, M. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience* (cit. on pp. 25, 42).

Van Lange, P. A. M., De Bruin, E. M. N., Otten, W., & Joireman, J. A. (1997). Development of prosocial, individualistic, and competitive orientations: Theory and preliminary evidence [Place: US Publisher: American Psychological Association]. *Journal of Personality and Social Psychology*, *73*(4), 733–746. https://doi.org/10.1037/0022-3514.73.4.733 (cit. on p. 44)

Vansteenkiste, M., Simons, J., Lens, W., Sheldon, K. M., & Deci, E. L. (2004). Motivating Learning, Performance, and Persistence: The Synergistic Effects of Intrinsic Goal Contents and Autonomy-Supportive Contexts [Place: US Publisher: American Psychological Association]. *Journal of Personality and Social Psychology*, *87*(2), 246–260. https://doi.org/10.1037/0022-3514.87.2.246 (cit. on p. 45)

Velmans, M. (2014). The evolution of consciousness. In D. Canter & T. D. A (Eds.), *Biologising the social sciences*. Routledge. (Cit. on p. 42).

Wiseman, T., & Yilankaya, O. (2001). Cooperation, Secret Handshakes, and Imitation in the Prisoners' Dilemma. *Games and Economic Behavior*, *37*(1), 216–242. https://doi.org/10.1006/game.2000.0836 (cit. on p. 5)