Prediction in Evolutionary Ethics: Towards Normative Autonomous Agents

SODI KROEHLER, The University of Pittsburgh

Perhaps the most constant paradigm in the realm of ethics has been that of constant disagreement. However, in the realm of autonomous agents, as well as other use cases, a great benefit would come from establishing a "shared base" of ethical theories - i.e. some sort of commonality between all the multitude of ethical beliefs. In this paper, I set out to explore what such a shared base might be. First I look through the most common ethical theories, and compare them with the ethical beliefs I got from interviews with several people from my social circle. Finding issues with each, I pull elements from sociobiology and Gauthier's contractarianism and attempt to situate reducing prediction as a possible shared base of ethics.

ACM Reference Format:
Sodi Kroehler. 2024. Prediction in Evolutionary Ethics: Towards Normative Autonomous Agents. 1, 1 (October 2024), 21 pages. https://doi.org/

1 INTRODUCTION

Many ethical issues in the computing space, as well as others beyond it, face difficulty in the fact that the ethical discipline is widely varied and rarely approaches a consensus on any matter. While some of this may stem from the viscous nature of the philosophical field, it also due to the nature of the issue itself. Ethics is not a primarily empirical discipline, and most ethical philosophers have explicitly avoided arguing along those lines. As such, it is difficult to meaningfully discuss algorithmic ethics, as ethical beliefs vary by person and possibly occasionally even in the same person and thus resist being generalized.

A principle issue in this domain is that of autonomous planning agents. Previous approaches (e.g. BDI [25]) have implemented a set of situational norms that could be queried to determine the ethical norms of the user. For many of quantifiable norms - such as which side of the road to drive on for an autonomous car - this is a satisfactory approach, yet as we will see (and as might be expected) this approach has difficulty in estimating more vague ethical beliefs. A value could thus be provided by some kind of "shared base" in the "ethical infrastructure". If all ethical theories arose from a single origin, or at least held some thing in common, this base could be used as a common standard and would be applicable to all users. These autonomous agents might then be able to build up from this to at least satisfactory models of the user's ethical beliefs.

To find such a thing, I first begin by surveying common meta-ethical theories, and comparing them to data found from several interviews, with the goal of determining if these ethical theories could serve as such a "shared base". Given the background of autonomous agents, I do this by staging the interview to attempt to extract at least one ethical norm, and for the sake of argument attempt to extract the "ethical cornerstone" of the participant - i.e. the self-identified most central difference between right and wrong in the participant. Then I find discrepancies between the meta-ethical

Author's address: Sodi Kroehler, The University of Pittsburgh, sek188@pitt.edu.

Manuscript submitted to ACM

theories surveyed and these identified ethical cornerstones and find, perhaps unsurprisingly, that these theories could not constitute such a shared base.

Following this, I note that one common element of all these ethical theories lies - somewhat *ipso facto* - in their assurance of the *ought*. I attempt to use this fact as a shared base, building off the work of [31] by mixing group selection theory from sociobiology with Gauthier's contractarianism to arrive at a expressivist but anti-realist meta-ethical theory, using reducing prediction as an ethical value instead of trust. I then show that such a theory would satisfactorily explain both the ethical cornerstones I identified in my interviews as well as the the common ethical theories that I surveyed from the philosophical discipline. Finally, it also shows promise as a much better candidate as a basis for autonomous agent normative reasoning.

2 LITERATURE REVIEW

I will now look at the state of literature in this enterprise. Since this is a multi-disciplinary study, I begin by surveying the state of the field in autonomous agent planning, and then move on to exploring the philosophical underpinnings of those approaches.

Ethical studies in algorithms are widely studied, with the three largest subdomains being privacy, mitigating bias, and ensuring explain-ability [12]. A smaller and lesser studied area regards the ethics implicit in the interactions between of autonomous agents and humans. This area is also interesting since, unlike the other subdomains, it entails a much closer look at the philosophical underpinnings of ethics proper.

One problem commonly surveyed is that of implementing norms in these agents. [14] - a seminal work on the topic - defines these norms as what would "represent the means to achieve coordination among agents which are assumed to be able to comply with norms, to adopt new norms, and to obey the authorities of the system as an end". Note that these norms are not implicitly obscure, they could be inferring the correct distance from a human [13] or which laws are applicable in which cases [17].

Norms are usually split into three modalities, or "obligations, permissions, and prohibitions for the normative agent who is the addressee of the norm" [13]. Krishnamoorthy et al. represents them with a 8-tuple, where the first three elements correspond to the situations where the norm is applicable, the next two denote its activation and deactivation functions, and the last three correspond to the penalties for acting contrary to it. Oh et al., given their subject matter, reduce the modalities to just obligations and prohibitions, and represent them in a 3-tuple, corresponding to the modality, the context condition, and the action(s) demanded. [14] uses a much looser definition than the others, with a constrained 4-tuple corresponding to the normative goal, the addressees of that goal, the context where it is applicable, and the rewards and/or punishments for not following it, respectively. Modgil et al. is yet still looser, and this time uses a separate category for norm expiration, but is otherwise similar to Lopez y Lopez and Luck [18].

How these norms are populated is often not thoroughly detailed. Arkin has spent a great deal of work in detailing how agents with lethal capabilities should be managed. Additionally, in many cases laws, ethics codes, and other legal documents provide a sufficient foundation [17]. All of the approaches detailed above would work with any of these. Beyond that, another option might be divine commands, which Bringsjord and Taylor advocates and implements using a three-law system (somewhat similar to Asimov's "Three Laws for Robotics").

Note, however, that all of these implicitly encode a realist assumption [6]. In the case of laws and even social norms, this may be acceptable - after all there is in fact a correct side of the street to drive on and a correct distance (albeit a flexible one) between robot and human. However in more ethics-related cases, this is not as clear, and certainly not without its critics. Additionally, it is not clear how to be sure that the set of normative rules is complete - even if there

is a finite and knowable set of ethical norms, there could be one that is so rarely encountered that it hasn't been written anywhere and will not for the foreseeable future.

This presents a significant road block to exploring a more general normativity in these agents. Since it is not clear how to generalize from such norm specifications to a novel scenario, such agents may be seen as less robust. These agents must interact with various parties who may or may not all agree to the same set of moral truths (assuming a realist position), as well as parties who do not hold a realist view of ethics at all, and while conflict resolution methods are certainly in place [25] [30], these methods require that the ethical beliefs of each participant be clearly delineated. Thus, an inordinate degree of responsibility is placed on the developer, and trust is difficult to engender[31].

However, I could not find anything that showed that a realist stance was required in these situation. In fact, it would seem that an agent with a non-realist meta-ethical stance could still interact with realist parties easily, if the norms that the agent operates under are considered as concepts, which are shared by both the realist and the non-realist traditions [7]. This would be a significant advantage for the agent, and theoretically improve its robustness. Additionally, depending on the method used, it may be much better at accounting for previously unknown situations autonomously. This is because the problem, at least for the agent, is reframed from an elucidation of an ethical fact (which they may not even be able to do - e.g. in divine-command ethics) to a measurement of the ethical concepts of the parties it is interacting with.

However, the non-realist ethical field is heavily nuanced, and it is not clear which standpoint would be most effective in this enterprise. As such, I decided to build on the work of Tariverdi, since it is both recent and provided a decent starting point.

2.1 Sociobiology

The focus of Tariverdi on trust is in line with a non-realist, expressivist and naturalistic standpoint. It is separate from the Theory of Planned Behavior [2], as it does not require that the ethical motivations of the particular user are entirely inscrutable. Additionally, although he does not explore it heavily, it is line with a naturalistic explanation, if combined with Gauthier's contractarianism and sociobiology - but I will show this later.

2.1.1 Origins and Constituents. Sociobiology itself arose in the 1960s, as an attempt to extend evolutionary biology into the realm of sociology. However it is probably best known by the work of E. O. Wilson's in his 1975 book Sociobiology: The New Synthesis [16]. It has since been expanded on and introduced to several different sub-fields, with varying successes [9] [24] [20] [33]. A very meaningful discussion on both its history and content can be found in [34].

The principle assumption in sociobiology is that the same mechanisms that underlie phenotypical changes in species across time can be applied to the development of humans culture and behavior. It is - perhaps obviously - decidedly Darwinian, but it separates from it in several key areas.

Firstly, when moving from phenotype to sociological artifact, the method of generation expands from a singular device (namely simple random mutation) to random mutation coupled with autonomous, voluntary action. This is an entirely intrinsic change - we are describing beings that have free will (namely ourselves) and thus must account for this ability in any argument we have.

Secondly, when moving from phenotype to sociological artifact the method of selection expands from simple survival and reproduction. However, here it is not as clear as to what it expands into. Various sociobiological theories have different explanations.

Kin-selection theory states that behaviors and artifacts are retained and developed if they improve the survival and reproduction of those who share similar genetic makeup to them [10]. Thus, a behavior would be most valued if it contributed to the well-being of one's children or immediate family, slightly less if it contributed to one's extended family, and so on with one's country, and finally species.

Group selection theory looks at the mechanism of evolution - namely, generation followed by selection - as happening on a group level at the same time as it is happening on a personal level. On a personal level, a behavior is promoted if it improves either the survivability or the reproduction of the individual; on a sociological scale, a behavior is promoted if it improves the survivability or reproduction of the group. This theory seems to show clear evidence in the world of eusociality. [28]

Multi-level selection theory seems to accumulate all of these theories under one umbrella, simply stating that the evolving processes are continually happening at multiple levels. It is extremely difficult to argue against this paradigm altogether on a strictly biological level [34].

2.1.2 Applications to Meta-ethics. There has been repeated efforts to use sociobiology as a basis of ethics. Herbert Spencer [28] is likely the first to begin an ethical theory in harmony with evolution. He is widely accredited for coining the term "Social Darwinism".

Philosopher Michael Ruse is arguably the largest name in the enterprise [26] . However, there are a good many others involved in the larger ideological area [15]. Of special note is Peter Turchin's account of historical activities from a sociobiological lens [33], providing a compelling viewpoint for group selection theory.

2.1.3 The Natural Fallacy. Numerous criticisms have been levied against these theories, both for accuracy and for more normative reasons. The first and most principle accusation is that of the "naturalistic fallacy" [19] [5]. This argument relies on the assumption that the "ought" exists in some form alternate from the "is", and thus it is a fallacy to draw from "is" to "ought". This is separate from the realist/non-realist debate since, to them, even non-realist ethical theories should draw their obligations and permissions from something other than observed behavior.

The "naturalistic fallacy" was originally defined in British philosopher G. E. Moore's 1903 book Principia Ethica as the following: "To argue that a thing is good because it is 'natural' or bad because it is 'unnatural', in these common senses of the term, is therefore fallacious;" [29]. It is perhaps the most common answer given against naturalistic ethics [19], and is frequently thrown at all manner of ethical theories. It is unmistakably a form of Hume's "is/ought fallacy", but was rephrased to allow it to work together with realist theories.

It is not the place of this paper to make a meaningful contribution to this discussion. However, since this draws from a naturalist explanation in the form of shared trust, it is not exactly subject to this fallacy, and becomes more Kantian than it does descriptive.

2.1.4 Gauthier and Contractarianism. Gauthier deserves special mention, as in some ways he bridges traditional contractarianism with sociobiological perspectives. A principle and valuable thought experiment in Gauthier's work is that of a stag hunt. In this thought experiment, an early human community is faced with hunting food. Two available options are set out for them - either each can go an hunt a rabbit on their own, or they can work together to hunt the larger stag. If everyone hunts rabbits, everyone will eat, but the rabbit population will quickly dwindle and all of them will likely die. If some go and hunt the stag, while others hunt rabbit, the rabbit hunters will eat well and the stag hunters will starve. The optimal solution is for all members to hunt the stag, leaving the greatest amount of food for them all as well as the least intrusive to the environment.

In this thought experiment, he shows the power of the survival of the fittest as seen through the medium of planning. Since the individuals here have higher cognitive functions, the battle for survival is taken a step away from immediate death/reproductive rights as it is with lower cognitive animals. Additionally, as seen through this thought experiment, Gauthier shows that the individual benefits from treating his neighbors kindly, as he benefits more from working with them than he does without them.

2.2 Measuring Personal Moral Philosophies

Due to both the constant burden of avoiding the naturalist fallacy, as well as the nature of the field, it is substantially difficult to measure personal ethical theories. As such, and as is typical in many philosophical conversations, it is often preferred to eschew typical social study behaviors in favor of a more "pure" reasoning structure.

The groups that do attempt to measure ethical theories directly from populations usually take the form of surveys on particular stances or else querying responses to various ethical dilemmas.[22] [27] I did not find any that had attempted to extract ethical beliefs in the way that I had. [23] is unique in reasoning explicitly towards an evolutionary ethics theory, and is thus worthy of special note here.

3 METHODOLOGY

To demonstrate the issues with a realist normative framework as well as the value of non-realist one, I did a case study of observed ethical beliefs, by interviewing several participants on their ethical beliefs and then attempting to encode these beliefs in a machine-readable way. To reduce specificity and broaden applicability, I attempted to extract the most important ethical belief in each of them - i.e. some sort of "ethical cornerstone" that would apply in most if not all situations that they were in. More technically, I sought to create a normative tuple $< v, \alpha, \mu >$ where v is the deontic modality, α is the context condition, and μ is the normative demand; I also aim to set α to ϕ, and thus include all possible contexts for this participant[21].

I do understand that to do so comprehensively is impossible, as it would not be possible to demarcate a participant's ethical beliefs for an infinite number of circumstances. However, note that as of yet it is also impossible to definitively extract their ethical beliefs in any situation, as it is not even clear if the ethical beliefs that they report are actually what they follow (see [2] for details, although there are many other such studies). The hope is that through this process a common theme or trend is seen, which could possibly constitute a shared base of ethical thought. If this is true, or even if it is largely true, and if this shared base is expressible in a machine-readable format, then, in line with a naturalistic and expressionist ethical standpoint then, an autonomous agent that employed it would be significantly more robust.

3.1 Participants

In this study I recruited six participants from my social circle. There were four men and two women, ages 23 to 81.

All participants were raised in contemporary Christian households, and all but one continue to profess a Christian faith. This was intentional, to restrict the study to a singular subset of religious experience. However, participants ranged widely in their commitment level to religion. Participant E works as a construction worker, and was the only one who does not identify as Christian, although he made an effort to acknowledge the existence of higher beings of some sort. Participant C works in an animation studio, and tries to attend church on Sundays. Participant G is a retired United Methodist pastor and has been married to Participant J, a retired Sunday School teacher, for over 50 years. Participant K runs a business with her husband, where they attend church every Sunday and make an effort

to hire other Christians. Participant T currently lives full-time in an intentional community with strong ties to both Christian and Jewish faiths [32].

Due to the nature of the project, my selection criteria was pretty much defined by who would answer my emails. However, I did later make the choice to restrict my emails to Christian-only participants, as specified.

3.2 Data Collection

All interviews were done over Zoom, with both video and audio recorded. Interviews lasted from 22 to 43 minutes, with an average length of 32. Participants G and J were interviewed together, while the rest were interviewed alone.

Participants were asked to be interviewed in regards to a university project looking at religion and ethical beliefs, along with the effect of technology on both of these (if any). Although I slightly varied the questions and their wordings based on the participant's responses, most interviews followed this general pattern:

(1) Please give us the background on the role of your religion in your life (e.g. how long you have practiced, whether many others in your social circle also practice the same.)
(2) Would you immediately think more or less about a person by learning that they practice your religion?
    (a) Why?
    (b) What about a different religion?
(3) How do you feel technology has shaped the practice of that religion?
    (a) Could a person effectively engage in your religion exclusively through technology (video chat, online reading, etc)?
(4) What do you think is the primary difference between someone being in the "us" group, versus the "them" group.
    (a) Do you think this would ever (or has ever) change(d)?
    (b) Do you think everyone should share this perspective?
    (c) Do you think this would still hold if we were on another planet, in some fictional world, etc?

3.3 Data Analysis

I took notes during each interview, combining them with the transcripts before coding. I then did open coding of the transcripts using Excel, followed by axial coding to get common themes. Since I was the only coder, I repeated this process after a week, and then again a few days later. However, the codings for two interviews were mistakenly deleted, so those participants K and T were only coded once.

4 FINDINGS

4.1 Shared Themes

4.1.1 Defining Religion. Defining religion was an important aspect of all participant's responses except for J. Participant T emphasized religion "being re-linked to god" and frequently made a distinction between a true religion, which demanded total obedience, and other religions who allowed other priorities. Participants C, K, and G were much more open in their definitions. Participant C emphasized a difference between "religion" and "faith", with faith being the motivator and religion the outward aspects. While careful to maintain that the outward aspects were essential, he also acknowledged personal shortcomings in this department, and focused on faith as the central part of their identity. Participant K didn't focus as much on literal definitions, but the theme was still apparent throughout her responses. She made heavy use of a "tree" analogy, where "religion" functioned as the trunk and outward behavior constituted the

branches. She acknowledged that other religions come from the same trunk, and thus had harmonious elements, stating instead "what makes Christianity very different is that there is definitely a reciprocation there" [K1]. Participant E, the one participant who didn't identify as Christian saying "I would say I would be against any and all forms of organized religion". However, he quantified this by separating religion from spirituality in that the former had strong tenets of human authority while the latter was more open.

4.1.2 Judging. All participants showed uneasiness when being asked to judge others. Additionally, all maintained that one should be humble, and it was very easy for people to be wrong. All participants also labeled one or more things that were beyond this level of acceptance, calling it either "objectively wrong" or in K's case "something I'm very sensitive to".

4.1.3 Technology. The responses on the technology questions were usually answered quickly, and I found it difficult to elicit more depth on those topics. Additionally, I amended my research question slightly after beginning to work through the responses, and decided to move away from this aspect. Hence I won't detail the responses for this section here.

4.2 Individual Themes

While the participants identified many shared themes, each one argued to a central division between good and evil - i.e. an "ethical cornerstone" that was unique among the participants.

4.3 C

As a Christian, living like Jesus was continually brought to the forefront. However, he does understand that this is not shared by others. The codes of other religions - he identifies Jewish and Muslim faiths - aren't held at the same level as the Biblical ones, but are still held having greater value than those who have none. Having a code or a cohesiveness seems to be an important aspect to him overall - he identified someone who was "just chaos" as a ethical evil, but also added that such a person was probably not possible.

A frequent theme in his answers was Jesus/God as an *archetype*. That is to say, in so much as other religions and peoples did well, they did so as sort of a result of the workings of the Christian God. "Jesus is the Son of the God that created everybody, that created all beliefs. It's like a family tree, it goes out that way, I guess I would say". This causes him to exhibit sympathy and a resistance towards judgement of other religions, since in his mind the faith he has is not even just a precursor to them - it is "is inherent to existence".

Along this was a theme of pity, both for those of other religions but particularly atheists. He noted that he was friends with a lot of atheists, and did not think them to be unethical, but just mistaken. I did clarify that it wasn't the lack of a support group or other believers that he found sad, but "they are missing a certain truth that will actually give them more freedom than what they're initially assuming" .

Finally, C made an effort to separate himself from total moral relativity. It wasn't enough to simply act in accordance with one's own ethical standard in some cases. "If someone says like, 'I love being a pedophile and I think it's very good'...I think that there is objectively wrong" . He noted that ethical standards have changed over time - citing the example of the age of marriage changing over time - but stated that "it's not up to men what makes ethics, but it's up to something deeper, which I guess I would say is like archetype" .

4.4 E

E stands in stark contrast both to C and the rest of the participants. Importantly, he does not identify as a Christian, and thus his ethical standards do not make any religious claims. However, he did identify his religious upbringing as the cause for making him skeptical both of religion and furthermore all types of authority.

In response to question 4, E said "It's if you hold yourself to the same standard as you hold other people to ... kind of like 'do on the others as you've want done to yourself'." However, later he quantified this standpoint heavily, much more than the other participants.

Like C, E maintained a standard of ethics beyond cultural norms. He identified the case of a serial killer, saying "We can't overlook what they do just because they don't personally think they're doing something wrong" and "So I wouldn't consider him unethical. I would consider him damaged and not fit to be a part of society, but I wouldn't say that he's past, you know, he hasn't gone past his own ethics because that particular thing just doesn't exist to him".

Also, interestingly, E is the only one out of the participants to define a time when he *would* break his ethical norms. He states "From personal experience, I guess I'd say like, if I'm having a hard time paying the bills, or money's harder to come by, and I still need to feed my kids, I would be more likely to break my ethical boundaries to take care of myself and the people close to me".

Another unique idea was that of ethics as a societal tool. "I think in order to have a cohesive society, we do need rules that are generally agreed upon" he says, when asked how to coexist with people with alternate ethical theories. Again revisiting the example of the serial killer: "They just need to realize that if the rest of the world thinks a way different than they do, that they should think carefully before engaging in that". Here, the focus is much more on societal health than the pleasure or lack of pain of the people in that society. "And as long as the society is working well, then that tool is kind of built in. But if the society starts to fail, then the, so does the tool, I guess" .

In response to question 4c, E affirmed that these tools for societal health would likely remain, although he wasn't sure if that were true if the species were different.

4.5 G and J

Participants G and J wavered somewhat when asked about the primary difference between right and wrong. Although I varied the question a few different ways, they both primarily gave answers that spoke of withholding judgement and remaining objective.

G repeatedly identified curiosity as a very important good. He repeatedly spoke of biases or instances when he had dismissed the other - specific instances being Catholics and those with disabilities. In line with this, especially G emphasized not wanting to judge others : "about the time I want to really be complaining about somebody else's failures, something happens in my life and I think" ... "you really should be careful there"..." you screw up regularly".

G and J were the oldest of the participants, and incidentally were the most open about the ways their ethical beliefs have changed. G remarked, and J agreed, that they had previously been unwelcome to Catholics, and refrained from socializing with them, but had since come to a more accepting view.

J offered hurting children as an unquantified evil, at least twice. G agreed, but also added some markers of good behavior - giving to charity, being faithful to your spouse, and engaging publicly with your faith. Faith was obviously an integral part of their careers, and they traced much of their thinking and value systems to it. However, they also pointed to examples of ethics around them, noting that many religions have aspects of "care for the other" and "seeking

God's direction in their life". "It has humane and good order and healthy personality kind of roots" they said would be the characteristic of a well-functioning moral outlook.

4.6 K

Similar to C, the concept of Christianity's God as the root or source of other religions and even good itself was also shared by K, but not to the same extent. The trunk and branches analogy was a common refrain throughout the conversation, and although she didn't say so exactly, it seemed more like Christianity encompassed more of the "tree" than other religions, whereas with C Christianity formed the roots and the trunk. That is to say, K approached other religions as being similar in substance but equal and different in appearance, whereas C viewed other religions more as corrupted or incomplete versions, when compared to Christianity.

That being said, her personal relationship with religion was very important to her. She describes a shared camaraderie that she feels with other Christians, and notes that while she works with and is approving of those from other religions, she greatly enjoys working and being around other Christians. Like C, she also identifies a sadness or pity for those from other religions who do not have such a personal relationship with their faith.

She often used anecdotes or stories to answer her questions. For example, she described a traumatic early experience - a neighbor's child dying in their pool - as cauterizing her decision to pursue a career in undertaking. She noted the differences about how they responded to the tragedy versus how she would have. She noted that both families attempted to convert one anther, but there also expressed some assurance that it was still the same god in both religions.

A principle cornerstone for K was "growth". "it's more like ... what really matters? Do the leaves matter? Do the branch matter or does the trunk matter? And to me, the trunk and the roots are what. And more than just roots, like I think you need to have that active growth from the roots continually". She was averse to labeling people, maintaining that you never really knew the full story behind someone's actions. However, she gave nothing more as a signifier between good and bad other than this "growth".

K's only quantification was that of Satanists, which she described herself as being "very sensitive too". An interesting anecdote she shared in defense of this involved a young girl at a funeral who would rub a ring on her finger every time "God" was mentioned. K reports being impressed by the girl's devotion, and then feeling separation bordering on disgust upon learning it was Satanic.

4.7 T

Although it was also mentioned as a shared theme, the definition of religion seemed to be very important to T. He began the interview immediately by defining religion as "re-linking to god" and he revisited on it many times throughout. True religion had a clear result of its members participating wholly in this "re-linking" - being "a hundred percent committed". He noted this is different from most other Christians who "probably say they give a hundred percent, and that translates to one day a week at a church service or something". He would not entirely confirm that the only truly saved people were a member of his group, but I also did not push on this as it was not entirely relevant to this paper.

For him, religion was largely separate from ethics. "Good" was instilled in each person, on account of them being created by God, and they either chose to listen to this "voice" or else deny it. He drew evidence of this natural law mostly from the Bible, depicting the story of "the barbarians of Malta". In the story, a group of island settlers came upon the main character Paul, who had began to build a fire after suffering shipwreck, when he was bitten by a snake. The settlers proclaimed that "although he has escaped the sea, justice has not allowed him to live". According to T,

this showed that the natural law was written upon all humanity, and that it demanded a great many ethical standards, among which was capital punishment.

For T, laws are meant to approximate these moral standards. He maintained that the conscience was universal, but said it was "a quiet voice ... in the heart of each person", and thus laws were there to enforce this small voice. However, he also identified cases where the law would transgress this - specifically in the case of the requirement to wear a seat belt, which he said was more of a form of social control than it was enforcing a moral norm. However, he also didn't object to it, saying "I still wear my seat belt ... but it's actually limiting my own freedom to protect myself". There were other cases that he was not so accepting of - he identified cases where capital punishment was being repealed, and in the case of the country of Germany which had outlawed parental discipline. In the latter, he explicitly identified it as a case where the "laws are in opposition to the moral law".

He did not have a clear answer when asked about how one might arbitrate in situations where the constitution of the moral law was disagreed on. He he brought up an anecdote about a former pastor he had known that had cheated on his wife and justified it by saying "it was the will of God". T said the real decision would come on Judgement Day, but "I'm inclined to think that the man was committing adultery and is gonna have to pay a penalty for that ". This was somewhat at odds with his statement that the natural law was "universal", but he didn't address this disparity even when questioned about it.

T exhibited some difficulty in answering question 4c. He first gave a prophetic depiction of the end of time, where despite the Messiah returning and forming a perfect world, there will still be people who complain, and are "done away with" for not showing gratitude. When prompted further to identify if the moral codes he had specified would still hold true in a different reality, he said "I don't really know how to answer the question" and "it seems like a pointless, meaningless question". He followed this by saying that there are things people think are one way and that "When Judgement Day comes ... they woke up or reality hit". However, he didn't specify further as to how one might know this reality before that time.

T also included some closing remarks about the separation of the "spiritual" and "natural" realms. He noted that "the natural man cannot understand the ways of God". He also added that empirical (aka "natural") lines of questioning were always at times incomplete, and occasionally at odds with the "spiritual" understanding. He mentioned how Google (it seemed that this was more of a paraphrase for all internet knowledge-transfer mechanisms and not just the specific company, but I did not confirm this) contributed to a society where "truth" was only what was available online, and that "it's almost like such information doesn't exist like if something can't be found on Google, does it exist?"

4.8 Ethical Cornerstones

In line with the main purpose of this study, I tried to arrive at a single "ethical cornerstone" for each participant - i.e. the principal difference between right and wrong. I have summarized these cornerstones in the table below.

Participant	Cornerstone
C	Behave like the archetype of Jesus
E	Live and let live
G	Be curious, not judgemental
J	Avoid harm, especially to kids
K	Continually grow/be connected to the "trunk"
T	Listen to your conscience

This extraction process was entirely qualitative, and is done merely to make discussion easier. C argued along very similar lines as did K, and thus in the interest of obtaining some kind of "shared base of ethics" they could definitely be grouped under a better-worded (and possibly more general) cornerstone. However, in this case I was attempting to just draw out the main identified difference between good and wrong from the answers given.

C directly identifies the "archetype of Jesus" cornerstone, saying "it's not enough" just to act in accordance with one's own ethical code, "its up to something deeper, which I guess I would say is the archetype". Nowhere else in the interview did he seem to change this, or else say something out of accord with it.

E's cornerstone here I took a little more liberty with, as he often quantified his statements. He responded to Question 4 by saying "I would say the biggest thing is, like, it's not so much your personal ethics. It's if you hold yourself to the same standard as you hold other people to". However, as we discussed, he also brought up the element of the psychopath, which although he didn't label as unethical, he still disapproved of. Additionally, he also made comments when answer Question 1 that emphasized his respect for other religions but only if they didn't "interfere with my life at all". Thus, it seemed like individual freedom was key to him, above any ethical code. However, I could also have identified the "self-cohesion" as his cornerstone, as it was also very important to him.

G continually identified being curious, in almost every answer that he gave. It was what he urged others to be, it was what he set forth as a good. There was not much option for a different cornerstone. J often agreed with G, and since it was a group interview I wasn't able to get more of her thoughts. Hence, I picked her ethical cornerstone from the few things she had added.

K was perhaps the most difficult, as she identified a lot of themes. The "trunk" analogy was revisited the most, and so I picked that. She also identified wanting to "be known as a Christian" and "be known for her love of others" but that didn't seem to be as guiding of a difference between right and wrong. Most of the other things that she described with a normative label could be better grouped into what made her religion better rather than what made her ethics better. However, she was also my first interview, and I added question 4 after this interview, so I likely would have had better results if I had included this question in that interview.

T also shifted many times between describing "good religion" and describing ethics. However, when talking about ethics, he would almost always mention the conscience or some description of it - such as the "still, small voice". Additionally, he put a great deal of emphasis on the group he was in, and this group has also published a great many papers acknowledging the same, so I felt fairly confident in this choice.

5 DISCUSSION

As stated before the goal of this paper is to explore the so-called "infrastructure of ethics", looking for any way to describe a shared basis for human ethics, and as such form a more robust basis for an ethical reasoning autonomous agent. Such an enterprise cannot be executed exhaustively - there are many humans who have died who we cannot ask about their moral beliefs, and even asking everyone alive would be too expensive. However, since I am set to not determine an entire theory of ethics, but rather just the largest possible commonality - the greatest common ethical denominator - I need to find something more general. To that end, I note that the entire ethical community - and perhaps a great majority of humans - unilaterally assumes the existence of the "ought". Thus, ethical conversations are reduced to what the ought is, with maybe some conversations expanding to how the ought came to be. Thus, an examination of the ought itself seems to be a great starting point for such a "shared base" of ethics.

5.1 Assumptions

In this line of reasoning, I make the following assumptions of any candidate for a shared base:

- I assume that if there is any shared basis of morality, it should be able to explain the ethical beliefs of all humans.
- Such a shared basis needs a path of origin and a substance that is not opposed to analytic inquiry.

The first is necessary strictly by definition. If it is the case that there is a single common element among all ethical theories, it must be the case that it is common. Note that this makes no claim to the observed behavior of these people, but merely to what they profess to believe as the dividers between right and wrong. It would absolutely allow for someone who knew the right thing to do but did otherwise. However this "right thing knowledge" would need to be shared universally. Otherwise, instantiating an autonomous agent with this shared base would do nothing to help its robustness, and we are much better off with one of the other approaches detailed in the literature review, that at least has some concrete norms.

The second assumption is one required by the use case. If there was a shared base, but it lay outside of scientific inquiry, it is impossible to have any kind of algorithmic exploration on it. Thus, the only way to engage with it would be via human "interlocutors" who would need to "divine" or otherwise elucidate these norms, and then set them down in ways that an agent could follow - leading us directly back to the previous realist-based approaches. For this reason, I dismiss many Natural Law arguments that place God or some deity as the originator of ethics, as God lies outside of scientific inquiry, at least according to my research.

Note that, in line with the goals of the paper, there are no other assumptions (at least not that I am conscious of). Thus, this discussion will not assume realist or non-realist positions, nor the existence of any particular element of the "ethical beliefs of all humans" have, nor the fact that such beliefs do actually exist a priori.

5.2 A Survey of Ethics Proper

There is still a large field of philosophy currently dedicated to the discussion of ethics, and it would be bashful and imprudent to disregard them. Most of these theories also include some meta-ethical component of them, that might do well to serve as this "shared base". However, these theories necessarily disagree, with no clear way how to bring them into harmony. Thus, at most one could function as such a shared base, but it is as of yet unclear how to me to conclusively measure which one that should be. Nevertheless, I will undertake a small and abridged survey of them, using several mainstream ethical beliefs, and see if any could be given reason to supersede the others on the basis of the interviews which I undertook.

To run the risk of repetition, note that I am not measuring the ethical theories in comparison to what each of these participants did or did not do or said they would do - which would be measuring from standard to behavior and thus be a natural fallacy, but instead measuring from the theory to the held ethical belief - or standard to standard. If any of these theories can serve as a shared base, they should be held in common among all participants.

As a reminder, I set forth in the findings section the idea of an "ethical cornerstone", which is intended only as a linguistic tool to describe some approximation of maxim, truth, or other construct which the participant holds to be the primary difference between right and wrong. Furthermore, I have attempted to define these for each of the six participants, and included my reasoning for doing so, all in the Findings section.

Thus, I will now go through several popular meta-ethical theories (listed below) and examine these ethical cornerstones in the light of their meta-ethical bases, hoping for a high degree of shared congruity. I use the following lay definitions of the meta-ethical bases of each.

- Natural Law Here the meta-ethical basis is realist, where the normative facts originate by divine command.
- Deontologicalism Here the meta-ethical basis is again realist, but the normative facts originate from the nature of reason [11]. Principally, there is also only one normative fact - the categorical imperative.
- Utilitarianism Here the meta-ethical basis can be seen as either realist or non-realist. Either way, normative facts (or concepts if non-realist) arise from things that preserve pleasure and minimize pain.
- Consequentialism Here the meta-ethical basis is non-realist, with ethical facts (usually only laws) arising from agreement of those governed by them.

5.3 Comparisons with Ethical Theory Viewpoints

5.3.1 Natural Law. Natural Law was, by far, the closest thing to a shared base of all the ethical perspectives identified. C identified wanting to evaluate others by similarity with their own ethical code, but also frequently identified God as being the progenitor of all other ethical theories, thus this could easily be interpreted as a natural law understanding. That is to say, he approved of each person living by their own moral code only because those moral codes were in line with the shared Natural Law which everyone possesses and also which was invariably written by the God he worships. This also explains his "pity" for those who were not in his religion, as they were readers only of the work and not friends with the author.

While E didn't explicitly maintain that the spiritual beings that he assumes to exist had any meddling in the nature or creation of humans, this certainly did not excuse him from describing a personal ethical theory that was very much in line with the Natural Law theory. He also made numerous statements in line with this, saying directly, "well I think every person instinctively knows right from wrong" .

While G and J were primarily action-based in their ethical theories, the concept of curiosity for the other also seems to rely on a Natural Law assumption. G's continued refrain that everyone is a "child of God" seemed to indicate not just a metaphysical truth but also a meta-ethical stance that because of this, each person could be expected to live according to how this God intended. Of course, both he and J also identified cases where a person might transgress this, but both seem to see a Natural Law currently in place and held commonly among all humanity.

K was very close to C on this matter.

Natural law was probably the largest aspect of Participant T's responses. A person was good to the extent that he acted according to the "good inside him", and bad if he didn't. Extra research into T's group reveals they maintain three classifications of humanity - the "holy", the "righteous" and the "wicked", corresponding to members of their group, saints, and sinners respectively. While the individual judgement for who belongs in each group is reserved for the "Judgement Day", there is a undisputed understanding that the dividing line between the "righteous" and the "wicked" categories is conformance to the "Natural Law" [32].

We might thus be tempted to conclude the paper here, and take Natural Law to be our best and most promising avenue for a meta-ethical shared base, since almost everyone's statements can be interpreted as coming from this standpoint. However, doing so would obfuscate the origins of such a shared base as being essentially entirely coming from a God or deity, which remains (at least according to my knowledge) to be outside the realm of scientific inquiry. Furthermore, there is little we can do to more accurately determine what this Natural Law is already comprised of. If we all hold it equally, and all have the same chance of both obeying and/or disobeying it, it is not clear how to determine upper and lower bounds on conformity to it, from which we might be able to deduce its demands.

If we cannot meaningfully deduce how it is constructed, nor what it is accurately comprised of, it seems like Natural Law itself is also outside the realm of scientific inquiry, and thus violates Assumption 2 for the shared base.

5.3.2 Deontologicalism. To participant C, the main thing was living like Jesus, which is primarily observable. When asked, he did state that a person could behave like Jesus, a.k.a. be a good person, and not be a Christian. This leaves a lack of clarity as to whether the principle differences lies in the person's motives or their actions. However, C also identified that others may not feel this way, and in that case he would try to evaluate them based on their own moral code, with notable exceptions - such as pedophilia - that were "objectively wrong". This is much more difficult to reconcile with a deontological perspective. If the categorical imperative was actually the substance that participant C's ethical cornerstone was formed from, without his knowledge (since he did not identify reason as the origin of his beliefs), where would this respect for the ethical standards of other's have arisen? If acting only on maxims which could universalize was the origin of his ethics, how would he have maintained an ethical cornerstone like "behave like the archetype of Jesus" while understanding that others would not or even could not act accordingly?

E seems most in line with a deontological framework, and goes so far as to nearly quote the categorical imperative verbatim. However, he also breaks its demands as frequently as he accords with them. He holds that the one who does a deontological wrong may imagine themselves to still be abiding ethically. The existence of such a person is not in discordance with a deontological system, but his acceptance of this person is. The serial killer in his example is acting outside of the categorical imperative, and yet is described by him as ethical and yet just mistaken. Thus, in much the same way as C, a deontological approach does not sufficiently describe his ethical stance.

G's value on curiosity is a problem in the same vein. Why would he extend this blanket status of "God's child" if reason alone had compelled him to value curiosity. J's specific ethical evils could be attributed to the categorical imperative, but "reason respecting reason" that forms the basis of Kant's theory wouldn't end up in universal acceptance. Thus, a deontological approach explains J but not G.

The priority that participant T placed on natural law is incongruous with a deontological perspective. It could be seen as being congruous, if we take the "still small voice" to be reason dictating universalizeable commands. However, capital punishment and parental discipline - the two concrete examples he gave - fit very poorly with this framework. Additionally, he did confirm that the intention to do good did not result in an action being good, and thus the deontological focus on motive seems to not be integral in his reason. Thus, again, it doesn't seem like the deontological perspective can be used as a shared base for his ethical cornerstone.

Participant K's notes about growth and shared basis of ethics could be construed as a deontological statement, in that "growth" was in effect the person's desire to do good. However, with her focus on the "trunk" and Christianity forming such a large part of the trunk, this method of explanation does not fit very well. At the very least, if the demands of reason could constitute a shared base for her ethical theories, why would it not be referenced. What would be the method in which religion overtook her reason? In a new situation, wouldn't she be more likely to reason along religious lines than reason lines?

The Kantian approach, then, does not seem to easily explain the ethical beliefs of any except for J. For G, it could explain but only condemning him for valuing those without a good will. For E and T it is actively broken, as both of these condemn the one who lives in contrast to the categorical imperative, but maintain that such a person would be thinking they were doing the right thing. In any case, although it does have a method of origin (namely reason) and a substance (namely whatever the categorical imperative demands) that are not opposed to analytical discussion and thus meets Assumption 2, it does not meet Assumption 1 and thus must be excluded from our candidates for a shared base.

5.3.3 Utilitarianism. All Christian participants (all participants except E) did not reason along utilitarian lines. Many of the things that were given as examples of good behavior - such as spousal faithfulness for T, G, and J, love for the

other for K and C - are of value in the utilitarian concept as well. However, since our conversation here is to determine a shared foundation for these theories, utilitarian values are not mentioned and the reasoning provided for their ethical cornerstones do not follow the lines of utilitarian reasoning.

The exception is participant E. While his cornerstone is much closer to libertarianism than it is to utilitarianism, there are many similarities. His reasoning often follows the lines of the golden rule, coupled with a sort of disregard for what others do so long as it doesn't interfere with him or his family. This focus on those around him in his ethical beliefs is still very incongruous with the utilitarian perspective. It may not seem so in that it is still important to him to promote pleasure and minimize pain in those whom he loves, but since he does not attribute a generalization to it, it is safe to say he would not engage in any kind of moral calculus at that time. He did identify his choosing of protecting or caring for his family over the fate of others as an unethical state, but yet even in his attribution of what was good there was no mention of generalization. Thus, it seems much more likely to conclude that he labeled these actions as unethical not because they held lower utilitarian value but instead because he had labeled those actions as such qualitatively, or at least using some other metric.

Thus, while we see utilitarian marks in the ethical cornerstones found, none of the participants could be labeled as reasoning along this lines. It does meet Assumption 2 in this case, but fails Assumption 1 and so it too is disregarded.

A critic might disagree on the grounds that just because the participants did not identify utilitarian ideas this does not mean they do not hold them. This is obviously possible, but if this is the case, it would then violate Assumption 2b, as there is as of yet no way to describe how these identified ethical cornerstones grew from utilitarian roots to the point that they make no mention of it or its premises. Thus, I still disregard this theory as a candidate.

5.3.4 Contractarianism. Participant E is also the closest to a contractarian perspective. In fact, this seems to be a closer motivation to his ethical cornerstone than does utilitarianism.

The other participants do not accord with a contractarian perspective, as they did not with a utilitarian. Although further data would obviously be needed, it does seem plausible that there is some negative relationship between a Christian outlook and a utilitarian or contractarian viewpoint. This would make sense given the belief structure - if God made humans, he would invariably have more of a say in what would be "proper operating behavior" than we would, and thus ethical models that rely on humans as primary elements don't match well. [1] Nonetheless, contractarian approaches fail Assumption 1 and thus must also be dropped from consideration.

Thus, all these common ethical theories seem to have issues explaining each of the ethical cornerstones of the participants, or else have issues functioning as this sort of shared base. If we were to program an autonomous agent with a library full of normative rules, it would seem that we would need an individual, context-free row for each of these participants, in order for it to handle unknown cases. However, we have no assurance that even these participants are entirely ethical, or that what they have identified as the main "ethical cornerstone" is actually what they believe, or finally that I have done a decent job at identifying this ethical cornerstone. Other attempts could surely be made, but with what scale would they be measured against mine? Even if they did, what should the agent do when faced with a decision that involves various participants with contradictory ethical beliefs?

[1] Deontological theories are still based in these, but I wonder if they show more similarity since they come from desire which is closer to the religious sphere than actions are?

5.4 Prediction as a Meta-Ethics

We thus have seen a set of similarities and differences between the ethical theories of the participants and pre-defined ethical theories. Natural Law was adequate as a shared base, but lacked clarity. Deontological theories seemed easier to attribute as the base of many, but was expressly violated in statements of acceptance for the other. Utilitarianism failed to account for most of the more central elements. Finally contractarianism explained the least.

In line with our initial line of reasoning then, we'd need to set forth a new meta-ethical standard. It'd need to be one that would encompass, or at least explain the other meta-ethical theories, to avoid building anew outside the foundations of history. It would need to both explain and even have intrinsic to it the concept of difference, since even among our six participants, ethical theories were set forth that are distinctly different from one another. As a "shared base", it would need to allow for this without breaking. Finally, it would need to avoid falling into the "naturalistic fallacy" - i.e. it would need to situate itself as somehow "above" observed actions and behaviors.

As promised in the introduction, however, I'll first take a closer look at the concept of the natural fallacy. It requires that the "is" be separate from the "ought" but does not describe where each comes from. The "is" is plainly obvious - it is whatever is found by experience or study. But the "ought" is a feeling itself - it has no necessary basis outside of that. What would be the problem with treating it exactly as such, *and nothing further*? In meta-ethical terms, let's move from a realist to a non-realist approach[7].

To even set out on such a path, we'd have to explain its existence. What would cause a person to acknowledge an ought - to believe that there is a standard for right and wrong? If this reason is evolutionary, how could this have arisen in early man? If it is cognitive action or "reason" shouldn't we see observed arguments that resembled contractarianism (as Gauthier explains), or deontologicalism? Perhaps there is another explanation that blends these two together, but I haven't found it yet. Instead, since I've already disqualified contractarianism and deontologicalism, I'll go forward with an evolutionary approach. [2]

First, note that group-based behavior is occasionally more individually beneficial than singular behavior, and that groups of individuals that work better together are likely to outperform groups that do not. Groups that work together are likely expected to have strong social norms to empower this, which might naively show promise of engendering normative belief from an evolutionary process. However we can also note from the sociobiological discipline that actions that benefit the group likely do not benefit the individual [34]. Thus, with simple evolutionary theory, individuals that behaved ethically would have benefited their group, but suffered themselves, and thus failed to reproduce and such traits along with them.

Introducing sociobiology, and in particular looking closer at this group selection theory, we might notice that ethics and norms are not strictly passed genetically. Thus, if we think of groups as a sort of individuals in their own right, this meta-ethical stance would have groups that behaved more ethically surviving and thus growing further by conquering groups who did not. However, such a theory still has a great number of problems. There are numerous examples in history of groups conquering others in which it would be difficult to attribute a greater ethical value to the conqueror [33]. Humans have also not evolved like ants or bees, in which individuals are unable to survive on their own and spend their time entirely on the benefit of the group[34]. Occasionally (and sometimes it seems often) the individuals who

[2]To be complete, there is obviously possibility that there are other possible origins - maybe deities or extraterrestrial forces or a host of other things. I can argue against most common divine command theories, since they usually are expressed in a way that means that ethical truths aren't scrutable enough to pass on entirely to machines, thus violating Assumption 2. But I can't argue against a position that a deity or something gave humans ethical reasoning - or even some form of the "shared base" that I am trying to find. I don't know even know to argue against that, and if it was true I don't know how to find it, but if an argument could follow evolutionary forces, it seems to me that Occam's razor would demand adhesion to it. However I admit this is a philosophical weak point in this paper.

experience the greatest in-group success are also labeled as being the most unethical. Thus, to use an evolutionary basis for our shared normative basis, we need something further.

5.4.1 The Development of Norms. As we have already identified, Gauthier lays a foundation for a correlation between evolution and ethics, through the medium of planning. I propose to use this as the missing link. Actions that benefit groups could only be undertaken if there was some assumption that the group would continue to stay intact. As with the stag hunt problem, you would only participate in the hunt if you were certain others would to. However, rather than illustrating the need for communal behavior, let's instead focus on the now-arising need for social prediction. Much like a prisoners dilemma, you now needed to predict the action of the others in your group, to know whether it was better to be selfish or altruistic. In fact, it has already been explored how this might have given rise to gossiping, or stories, or groups formed out of one's kin - all these things made prediction a little easier [1]. Even our sayings - "When in Rome, do as the Romans" - there is so much focus in life in reducing variation to understandable, predictable avenues - a.k.a norms. But it was still a difficult process, and was implicitly prone to error.

Again, norm creation by these methods is all but understood [8] [1]. Thus, why would it not be surprisingly to conclude that individuals that were better at predicting were thus more evolutionarily benefited? If this is true, would it not be the same for those who were easier to predict? Would not both provide a similar evolutionary benefit to the group? If you were a poor hunter, but you were enthusiastically showing up to every hunt, wouldn't you provide a value to the group strictly because they had to worry about you less? Granted, there are limits, and the better hunters probably are more valuable, but the paradigm remains - lower the work that others had to do to predict your actions made both your life easier and improved the health of the group.

To be fair, this line of reasoning is not altogether new, and contractarianism itself uses something similar as it's basis. However, it jumps immediately to the establishment of laws and governments as manual ways to arbitrate this ease of prediction. But these only came to be much later, comparatively, in the human development process. Besides, for our use case, we already have established that we can't use contractarianism since it didn't explain the respondent's stances. More specifically, if contractarianism is responsible for ethical norms as well as laws and social norms, what is the process for "converting" a norm to a ethical norm?

5.4.2 Mitigating Prediction. So, we have already decided to treat the ought as a form of norm (since we are espousing a non-realist stance), so there must be some system by which the ought is converted to the norm. And, since we're not accepting contractarianism, this is separate from the process by which a norm becomes a law. In contractarianism, a norm becomes a law when an agreement is made to "codify" the norms into a law - a miniature example being the mandated agreement to all hunt the stag. This could possibly be extended to norms as well, as sort of an intermediate step, except that our user studies show a separation of ethics and law as two separate, and sometimes competing, elements - rather than one as the continuation of the other.

Taking this background of minimizing prediction more forefront, however, would give a much more concrete separation between the two. An agreement to participate in the stag hunt would have to be made before every hunt, until of course it was codified into "The Rules of Stag Hunting". But then there would be other rules, like "The Rules of Soup Crafting", and the "Rules of Not Killing People". Then, we could imagine someone making the soup a different day one day, and the realization that it was quite alright to do so now showed that there needed to be a difference between "Stag Hunting" rules and "Soup Crafting" rules, and would thus cause an investigation into this difference.

This is where I push the prediction paradigm beyond what has already been established. Easing prediction could be hypothesized to be often done through self-reflection, i.e. what another will do is governed by what you might do. But

self-prediction is not very discriminative or robust, since each person is already seeing themselves as autonomous and thus could see themselves doing a wide variety of things. Thus, social norms occupy a place in the mind of "things I wouldn't do but I could see myself doing" and oughts take on the "things I would never do". As one encounters new behaviors, they continually would place these into one of these groups, with the oughts eventually being communicated to others. Eventually, these shared "theories of ought" would "wrap around" on themselves, being taken as guiding lights to the individual - standards that they as well as everyone else must follow. In other words, Cain was a stone-thrower pioneer, and then became a murderer once he asked himself what he had done.

It would come as no surprise that these oughts would be in line with things that were either good for the group or at least good for the species. Ethical standards such as pedophilia and hurting children would be perfectly understandable, since the primal "survival of the fittest" evolutionary constraints still remain in effect, and in this case don't conflict with group health. Because of the nature of reason, it would be almost a ethical norm in itself to believe in ethical truths, since to not made you so difficult to predict. If you didn't profess to believe in norms, you could be accepted as "someone who makes soup strangely", but if you didn't believe in these "oughts" - these "super-norms" - you had no agreements in place to limit your behavior and thus could be expected to do anything, thus severely raising the potential things to worry about involving you.

5.5 Mitigating Prediction as an Ethical Theory

If we take this as a shared base - meaning not only did the desire to reduce predictive energy create ethics, but is also the instigating force behind their adoption in each person, we should see evidence of this in the interviews.

It would explain easily why C, E and K placed so much emphasis on "self-cohesion", as the existence of some moral code was the most important to them. Having no code means that there is nothing restraining them from bad actions, which means that bad actions can't be ruled out immediately and thus the work of prediction is increased. Such a person could have a life full of actions which these participants would label as good, and yet likely there would still be unease since there is no immediate restraint on what is considered "bad". All these participants knew and accepted the fact that some people would have ethical codes that would enable them to do unethical actions, but that was somehow "better" than someone who could not be predicted at all - "chaos" as C said.

Additionally, as we discussed in the last section, it would explain the shared acceptance of the Natural Law theory, as well as the inherent "muddiness" surrounding its actual contents. We as humans would thus believe there to be one central source of ought simply because to not do so is so much more difficult. To attribute this thing which was always "observed" but never detailed would be a perfect candidate to attribute to a god. Thus, it would explain why so many of the participants attributed ethics to the workings of God - it is much easier to think of oneself as a child of this Creator than it is to imagine oneself as both the author and the subject of a giant universal-wide set of rules.

Reducing prediction would also very easily explain the deontological perspective, since that was - in some manner - its purpose all along. A society of people that all acted only for the maxim which could be extended to everyone would face as little predictive effort as possible.

It would easily explain the creation of the utilitarian theory, as prediction is only necessary in order to survive as a group. The group was only created to ensure pleasure and survival for its members. Thus reducing prediction is ultimately advancing a utilitarian ideal to begin with, but yet more robust and with the ability to explain more ethical standings.

Finally, it is nearly a brainchild of contractarianism, and thus can be seen as extending it more "fuzzily" to ethics than did Gauthier.

5.6 Implications

If this idea of reducing prediction then continues to work as a "shared base" of ethical norms, it would be much easier to encode this into an ethical-reasoning algorithm. As I stated earlier, I hoped to create a single normative tuple $< v, \alpha, \mu >$ where v is the deontic modality, α is the context condition, and μ is the normative demand. In this case, reducing prediction is advanced as a way to explain all three deontic modalities, and thus the v element could be removed. As I stated earlier, I hoped to let $\alpha = \phi$, and thus be applicable in all context, which again is true in the advanced theory. Finally, there is a single normative demand - reduce predictive action.

This is very analogous to Tariverdi's approach, since minimizing prediction is very similar to improving trust. However, it does differ from it substantially, since trust is more of the result in the minimizing prediction theory, rather than the origin.

It should also be noted that implementing this theory directly may work poorly, even on a theoretic sense. If this theory does hold water, it necessarily demands both oughts and the failure to examine those oughts, or at least to hold them above the level of norms. Thus, an autonomous agent that treated them in a continuous fashion - i.e. minimization problem involving predictive work - would likely produce results that are inconsistent with user's preference. In other words, even if we have a suitable shared base, it doesn't seem like building back up from this base will result in the ethical beliefs that we observed.

Additionally, it is not clear how to measure the effort that goes into predictive action, in order to stage such a minimization problem. However, with the minimizing prediction theory, the *entire problem* does become scrutable. Normative elements might still need to be encoded for various users, and there is expected to be a high degree of overlap - a good example would be the mandate to not hurt kids which likely all the participants would agree to. However, in the case of new situations, the course of action that showed the least bit of variation from the courses of actions which the most people were most likely to label as good would be the most ethical choice, at least until further discussion. An agent may even be able to help the human participant to be more ethical, by learning what actions most often prompted guilt, and seeking to prevent these actions - assuming of course that the feelings of guilt also arose from the individual failing to minimize their own self-prediction. When encountering a new participant, it could largely be assumed that they would follow the ethical mandates of those in their group (where group here can mean any kind of social group), with customization the more the user is exposed to both their actions and their professed beliefs.

5.7 Limitations

Of course, previous ethical theories have not tried to explain what is but instead argue for what should be. Thus, it is not the aim of this paper to critique them in any way, but rather to discuss only what a shared base could be.

That being said, although I have attempted to I know that this paper does not do sufficient justice to the ethical field, which it deals with heavily. Although I made many efforts to restrain my line of argument to keep only in the domain which I have enough knowledge in, I am sure that many ethical philosophers would find numerous opportunities for additions, if not widespread corrections.

Furthermore, it is easily observable that the experiments done in this paper are not nearly sufficient to prove these conclusions in any meaningful sort of way, nor even to suggest they are plausible. After all, even participant selection was highly biased, and there remains a non-zero possibility that every single person that I interviewed was unethical. This could be the case no matter how many people I surveyed. The purpose of this paper is merely to introduce a set of questioning into a possible shared base of ethics, and to introduce reducing prediction as a possible candidate. If

this theory can successfully account for the six participants given, there is some chance that it might also be useful for more, at least when compared to the others that could not even do that.

5.7.1 Ethical Considerations. While it does seem slightly odd to include a ethical considerations sections in a primarily ethical paper, it is nonetheless necessary. I feel it would be remiss not to acknowledge the false theories that have sometimes arisen on sociobiological foundations, such as specisim or naive group theory. I do not seek to contribute to these theories at all nor grant them any leniency nor veracity.

Additionally, while advocating a theory that brings about the ought from naturalistic means, I don't mean to advocate the freedom of any of us to disobey these oughts. Offering an explanation for why we feel the way that we do shouldn't give us any reason to stop feeling like that.

6 CONCLUSION

In this paper I have attempted to explore a basis for normativity in autonomous agents from a non-realist perspective. I did this with the goal of finding some "shared base" of ethical thought that might help the agent determine an ethical response to a novel context that had no ethical norms had been defined for it yet. In the pursuit of this, I conducted several interviews, attempted to deduce the "ethical cornerstones" of the participants, but found that most common ethical theories could not account satisfactorily even for the responses of the participants. As such, I combine several other lines of reasoning, from contractarianism and sociobiology, and advance a theory of reducing prediction as a possible shared base. I find that it explains the found ethical cornerstones to a much more satisfactory degree, and furthermore that it explains the existence of the other ethical theories which I surveyed. As such, it seems to be a promising avenue towards future algorithmic ethical reasoning.

REFERENCES

[1] [n.d.]. Gossip in Evolutionary Perspective - R. I. M. Dunbar, 2004. https://journals.sagepub.com/doi/full/10.1037/1089-2680.8.2.100?casa_token=fU1mCf0vK-gAAAAA:qRBK7RsIAYQxh6s-5p5seiYaRZvTqAUyj2JJacRYMCM3D1pOIHctBAcWYsl01b9c51UknELdTLFN0w

[2] Icek Ajzen. 1991. The theory of planned behavior. *Organizational behavior and human decision processes* 50, 2 (1991), 179–211.

[3] Ronald Arkin. 2009. *Governing lethal behavior in autonomous robots.* Chapman and Hall/CRC.

[4] Selmer Bringsjord and Joshua Taylor. 2012. Introducing divine-command robot ethics. *Robot ethics: the ethical and social implication of robotics* (2012), 85–108.

[5] Donald T. Campbell. 1979. Comments on the sociobiology of ethics and moralizing. *Behavioral Science* 24, 1 (1979), 37–45. https://doi.org/10.1002/bs.3830240106 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bs.3830240106.

[6] Mark Coeckelbergh. 2011. Is ethics of robotics about robots? Philosophy of robotics beyond realism and individualism. *Law, Innovation and Technology* 3, 2 (2011), 241–250.

[7] David Copp and Justin Morton. 2022. Normativity in Metaethics. In *The Stanford Encyclopedia of Philosophy* (fall 2022 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2022/entries/normativity-metaethics/

[8] David Gauthier. 1987. *Morals by agreement.* clarendon Press.

[9] Margaret Gruter. 1977. Law in Sociobiological Perspective. *Florida State University Law Review* 5, 2 (1977), 181–218. https://heinonline.org/HOL/P?h=hein.journals/flsulr5&i=191

[10] William Donald Hamilton. 1996. *Narrow Roads of Gene Land: Volume 1: Evolution of Social Behaviour.* Spektrum Academic Publishers. Google-Books-ID: yIk1sCPKrYkC.

[11] Immanuel Kant and Jerome B Schneewind. 2002. *Groundwork for the Metaphysics of Morals.* Yale University Press.

[12] Michael Kearns. 2020. The Ethical Algorithm. Book Title: The Ethical Algorithm Edition: 1st edition.

[13] Vigneshram Krishnamoorthy, Wenhao Luo, Michael Lewis, and Katia Sycara. 2018. A Computational Framework for Integrating Task Planning and Norm Aware Reasoning for Social Robots. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).* 282–287. https://doi.org/10.1109/ROMAN.2018.8525577 ISSN: 1944-9437.

[14] F. Lopez y Lopez and M. Luck. 2003. Modelling norms for autonomous agents. In *Proceedings of the Fourth Mexican International Conference on Computer Science, 2003. ENC 2003.* 238–245. https://doi.org/10.1109/ENC.2003.1232900

[15] Jennie Louise. 2009. I won't do it! Self-prediction, moral obligation and moral deliberation. *Philosophical Studies* 146, 3 (Dec. 2009), 327–348. https://doi.org/10.1007/s11098-008-9258-5

[16] Marjorie C Meehan. 1975. Sociobiology: The new synthesis. *JAMA* 233, 9 (1975), 1006–1006.

[17] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society* 3, 2 (Dec. 2016), 205395171667967. https://doi.org/10.1177/2053951716679679

[18] Sanjay Modgil, Noura Faci, Felipe Meneguzzi, Nir Oren, Simon Miles, and Michael Luck. 2009. A framework for monitoring agent-based normative systems. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*. Citeseer, 153–160.

[19] Julio Muñoz-Rubio. [n.d.]. SOCIOBIOLOGY AND THE ORIGINS OF ETHICS. ([n. d.]).

[20] Carey D. Nadell, Joao B. Xavier, and Kevin R. Foster. 2009. The sociobiology of biofilms. *FEMS Microbiology Reviews* 33, 1 (Jan. 2009), 206–224. https://doi.org/10.1111/j.1574-6976.2008.00150.x

[21] Jean Oh, Felipe Meneguzzi, Katia Sycara, and Timothy J. Norman. 2013. Prognostic normative reasoning. *Engineering Applications of Artificial Intelligence* 26, 2 (Feb. 2013), 863–872. https://doi.org/10.1016/j.engappai.2012.12.006

[22] Yue Pan and John R. Sparks. 2012. Predictors, consequence, and measurement of ethical judgments: Review and meta-analysis. *Journal of Business Research* 65, 1 (Jan. 2012), 84–91. https://doi.org/10.1016/j.jbusres.2011.02.002

[23] Lewis Petrinovich, Patricia O'Neill, and Matthew Jorgensen. 1993. An empirical study of moral intuitions: Toward an evolutionary ethics. *Journal of Personality and Social Psychology* 64, 3 (March 1993), 467–478. https://doi.org/10.1037/0022-3514.64.3.467 Num Pages: 467-478 Place: Washington, US Publisher: American Psychological Association (US).

[24] Tommaso Pizzari and Andy Gardner. 2012. The sociobiology of sex: inclusive fitness consequences of inter-sexual interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367, 1600 (Aug. 2012), 2314–2323. https://doi.org/10.1098/rstb.2011.0281 Publisher: Royal Society.

[25] Russell W. Robbins and William A. Wallace. 2007. Decision support for ethical problem solving: A multi-agent approach. *Decision Support Systems* 43, 4 (Aug. 2007), 1571–1587. https://doi.org/10.1016/j.dss.2006.03.003

[26] Michael. Ruse. 2009. *Philosophy after Darwin : classic and contemporary readings*. Princeton University Press, Princeton.

[27] Jatinder J. Singh, Scott J. Vitell, Jamal Al-Khatib, and Irvine Clark. 2007. The Role of Moral Intensity and Personal Moral Philosophies in the Ethical Decision Making of Marketers: A Cross-Cultural Comparison of China and the United States. *Journal of International Marketing* 15, 2 (June 2007), 86–112. https://doi.org/10.1509/jimk.15.2.86 Publisher: SAGE Publications Inc.

[28] Herbert Spencer. 1891. Progress: Its law and cause. (1891).

[29] Philip Stratton-Lake. 2006. G. E. Moore: Principia Ethica. In *Central Works of Philosophy v4* (4 ed.). Routledge. Num Pages: 18.

[30] M. Tamura. 2001. Multi-agent utility theory for ethical conflict resolution. In *2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat.No.01CH37236)*, Vol. 3. 1779–1782 vol.3. https://doi.org/10.1109/ICSMC.2001.973559 ISSN: 1062-922X.

[31] Abbas Tariverdi. 2024. Trust from Ethical Point of View: Exploring Dynamics Through Multiagent-Driven Cognitive Modeling. https://doi.org/10.48550/arXiv.2401.07255 arXiv:2401.07255 [cs, eess].

[32] Twelve Tribes. [n.d.]. What We Believe | Twelve Tribes. https://twelvetribes.org/beliefs

[33] Peter Turchin. 2018. Historical Dynamics : Why States Rise and Fall. (2018), 1–264. https://www.torrossa.com/en/resources/an/5573627 Publisher: Princeton University Press.

[34] David Sloan Wilson and Edward O. Wilson. 2007. RETHINKING THE THEORETICAL FOUNDATION OF SOCIOBIOLOGY. *The Quarterly Review of Biology* 82, 4 (Dec. 2007), 327–348. https://doi.org/10.1086/522809