Ethics in Artificial Intelligence
How Relativism is Still Relevant

By Loukas Piloidis, MA

Master’s Thesis June 2020
Södertörn University, Sweden
Abstract

This essay tries to demarcate and analyse Artificial Intelligence ethics. Going away from the traditional distinction in normative, meta, and applied ethics, a different split is executed, inspired by the three most prominent schools of thought: deontology, consequentialism, and Aristotelian virtue ethics. The reason behind this alternative approach is to connect all three schools back to ancient Greek philosophy. Having proven that the majority of arguments derive from some ancient Greek scholars (Socrates, Plato, Aristotle), a new voice is initiated into the discussion, Protagoras the Sophist. A big advocate of -the later conceived- humanism and relativism, Protagoras is used as a prism to examine a new ethical model that is based on the personalization of agents. In other words, even though theories of objectivity have overflown contemporary discussions of finding a robust ethical model for Artificial Intelligence agents, there is potential in a subjective model, personalised after each and every user.
Table of Contents

Table of Contents ................................................................. 3

Introduction ............................................................................. 5
  Background ......................................................................... 7
  Artificial Intelligence Definition ............................................. 9
  Ethics and AI ....................................................................... 10
  Introductory comments ....................................................... 13

Literature Review ..................................................................... 14
  Deontology .......................................................................... 15
    1. Transparency ............................................................... 15
    2. Artificial Intelligence omnipotence .................................. 16
    3. Artificial Intelligence Abstraction .................................... 17
    4. “Intelligence” ............................................................... 18
    5. Ethical Framework Deontological model ......................... 19
    6. Diversity is scarce ....................................................... 20
    7. Discrimination benefits ............................................... 21
  Critique of Deontological Ideas in AI .................................... 22
    Consequentialism ............................................................ 23
      1. Different consequentialist approaches ......................... 23
      2. Data Issues .................................................................. 24
      3. Hidden from public view ............................................. 27
      4. A Cheap Alternative .................................................. 28
      5. Biased AI .................................................................. 28
      6. Surveillance .................................................................. 29
      7. Corporatization of public Knowledge ............................ 30
      8. Universal Linguistic Need ............................................ 32
      9. Artificial Intelligence as a tool ...................................... 32
  Critique of Consequentialist ideas on AI .............................. 34
    Aristotelean Virtue Ethics .................................................. 34
    Critique of Aristotelian Virtue Ethics ................................... 37
  Solutions ............................................................................... 37

Common Roots in Ethical Frameworks .................................... 40
  Deontology .......................................................................... 42
  Consequentialism ............................................................... 42
  Final thoughts ...................................................................... 43

Discussion ................................................................................... 44
  Relativism over Objectivism .................................................. 45
It reminds me of Game of Thrones, where people keep being like, “We’re so busy fighting each other but the real thing we should all be focusing on is what’s coming from north of the wall.” We’re standing on our balance beam, squabbling about every possible issue on the beam and stressing out about all of these problems on the beam when there’s a good chance we’re about to get knocked off the beam.

And when that happens, none of these beam problems matter anymore. Depending on which side we’re knocked off onto, the problems will either all be easily solved or we won’t have problems anymore because dead people don’t have problems.

That’s why people who understand […] AI call it the last invention we’ll ever make—the last challenge we’ll ever face.
So let’s talk about it. (Urban, 2015)

Introduction

In this project, the aim is triple: a) I aim to delineate and analyze contemporary ethical frameworks concerning Artificial Intelligence; b) use ancient Greek philosophy to understand them; c) propose a new solution model based on another ancient Greek philosopher that has not yet been considered, Protagoras the sophist, who brings a rather subjectivist approach to a very objective, humanistic debate. The main issue that these frameworks try to tackle is that there is not a single robust ethical model that can shape how Artificial Intelligent agents will be ethically developed, used, and -in retrospect- judged.

Contemporary discussions about Artificial Intelligence are based on the grounds of three main ethical schools of thought, Deontology, Consequentialism, and Aristotelian Virtue Ethics.

In contrast to this circumscribed inclusivity where other schools of thought have been left out of the discussion, my proposed model is based on the theory of relativity, as initiated by Protagoras the sophist in 5th C BCE Athens. In other words, this thesis proposes that Artificial Intelligent agents should not follow normative ethical guidelines but instead, be personalized after each user and act the same way that their human agent would act in a certain circumstance. Using this subjectivist approach, most paradoxes and fallacies that have been so far generated under objectivistic theories will be resolved. Additionally, it brings an air of refreshment to a cluttered discussion from similar arguments that do not seem to promote the discussion into new levels. Although philosophy is notorious in having eternal questions
without clear answers, this case is different. Artificial Intelligence is a prominent discussion in our contemporary society. It would be safe to assume that as it gets normatized, it will be of increasing importance. Thus, in accordance with the theory of applied ethics, the problem of AI ethics is not just a mental quiz but a real problem with real consequences (Dittmer, 2020). Even if my proposed model is proved to be ineffective in terms of practical application, it will still generate novel argumentation, and promote a post-modern approach that has been so far suppressed in this context. Zygmunt Bauman argued quite convincingly on the importance of post-modern ethical theories, criticizing the over-reliance on utilitarianism by contemporary scholars (Bauman, 1993).

Since the collection and analysis of data about the specific subject do not require quantitative methods, the empirical research will be conducted through a qualitative and extensive literature analysis, which will be present in the section Literature Review. As a consequence, this project can be seen as a theoretical addition to the very divisive collection of qualitative body of research concerning Artificial Intelligence ethics. In other words, this project will examine core ethical debates concerning Artificial Intelligence, as well as the newest proposed solutions to overcome various philosophical and social obstacles that arise. The reader should keep in mind that an inclusion of the total body of research would be impossible for the size of this project, as well as the limited access that I have to, mainly, western literature.

Outlining the project’s basic sections, there are five main ones, some of them containing multiple subsections: Introduction and Background, Literature Review, Common Roots in Ethical Frameworks, Discussion, and Conclusion. In the introduction I hand out some general information about Artificial Intelligence, so that a basic level of understanding may be upheld by the reader. In the Literature Review I present the arguments that have dominated the scholarly debate concerning the social aspects of Artificial Intelligence ethics and categorize them into three main categories; Deontology, Consequentialism, Aristotelian Virtue Ethics. In the third section I continue the discussion from the Literature Review, declaring that all main ethical schools of thought derive from common roots, i.e. the classical ancient Greek philosophers, Socrates, Plato, and Aristotle. In the discussion, I propose that unfair discrimination has been held in Protagoras the sophist, another major philosopher of that era. Since contemporary proposals bear with them ancient doctrines, willingly or not, I do not see why prioritization of some ancient Greek philosophers should exist over others that might

---

1 I am aware that the philosophical branch of posthumanism would be interesting in this discussion, but unfortunately time and space considerations do not allow me the luxury of analysing it.
be just as relevant to this discussion. Finally, in the conclusion I set some limitations to this project, as well as I give some ideas about further research that could stem out of this project.

A final note before I delve deep into philosophy and technology is due. The “standard” demarcation of ethics is split into three parts, “normative”, “meta”, and “applied” (Fieser 2020). While this project seems to be concerned with mostly applied ethics, there are some instances where normative and meta come into the discussion. This is not due to a lack of focus. Rather, it is my belief that applied ethics are not very effective because they draw conclusions based on normative and meta ethics. In the words of James Fieser (2020), “By using the conceptual tools of metaethics and normative ethics, discussions in applied ethics try to resolve these controversial issues.” This is problematic because the last two branches have different, often conflicting issues. As a consequence, applied ethics are almost impossible to argue objectively.

**Background**

Artificial Intelligence (AI), for the 21st C, has been as influential in the business, academic, and media world, as space traveling has been for the previous one. Astronomy is still hugely invested in, of course, with academics and businessmen unfolding grand plans for the future. However, Artificial Intelligence has stolen momentum, as societies see a huge potential deriving from a concept that has been theorized before but has been highly neglected.

For businesses, Artificial Intelligence is not only virgin land, one that can turn out to be extremely profitable if correctly exploited, but also an opportunity for further advancement to their already established products/markets (Dirican, 2015). Artificial Intelligence can create a new market by completely altering the way this society functions, from autonomous cars to robot-assistants, from factory workers to security bodyguards, scientists, doctors, and athletes. This new market will be of insurmountable value to the industrialists that manage to conquer it, not only in terms of monetary profit, but also in terms of being in a position of power to “control” much of society’s functions.

It will also be of high value to them because of the way that said industrialists will shape society, much to their “mission” and “vision” -corporeal jargon that one encounters everyday now- as well as it will help them to gain certain positions of power (Skilton & Hovsepian, 2018). These positions, including political ones, can redefine the democratization of a contemporary society, and through lobbying and a well-polished marketing plan, be able to pass laws and/or legislation that will promote their agendas; and sometimes, those agendas are built on the exploitation of citizens. This is not mere speculation, as the reader will see, because in the next section, Literature Review, I am going to provide numerous examples of similar scenarios happening at the moment.

But there is another reason why businessmen are very much interested in the advancement of Artificial Intelligence. By incorporating said algorithms into their existing
products, services, or both, they can skyrocket their sales (Dirican, 2015). This will lead into bigger funding for their Research and Development departments, which will come up with even more sophisticated ways to dominate the market. While this is not necessarily a bad thing - all in all, if research funding increases and consumers get more top-notch technology that is indeed helpful, it is a win-win situation- one should bear in mind that this straightforward approach is rarely the case.

Most of the times, the business research is done on how to get ahead of the competition and, disregarding the utility of their outcome, it will only focus on getting an advantage. In other words, as we have seen quite recently, Big Tech companies (mainly Apple, Microsoft, Facebook, and Amazon) will do everything in their power to outperform their competitors, even if that means exploiting their consumers. In the Literature Review section, I am going to demonstrate multiple recent examples; but let me just point out here to the -arguably- more recent and biggest data mining scandal that have revealed the true colors of most of these corporations. Nicholas Confessore explained in his article “Cambridge Analytica and Facebook: The Scandal and the Fallout So Far” how the Cambridge Analytica scandal helped certain politicians to rise to power, and other political parties to reap significant benefits from a clear user exploitation (Confessore, 2018).

For academics, Artificial Intelligence is a double-edged sword; a potentiality that is both inherently intriguing in its essence, but deeply problematic in its core (Taylor & Wyatt, 1992). It is intriguing because rarely such a big field of study unfolds in one’s lifetime. Academia’s main purpose is to think, argue, depict, construct and deconstruct an idea or a field of study, to look at it from every different angle possible, in order to come to a collective conclusion if said idea or study field is worth incorporating in society. In this sense, this highly ambiguous, veiled notion that Artificial Intelligence is, is an opportunity of a lifetime for academics that want to add up to the collective bucket of societal knowledge for the advancement of contemporary society.

On the other hand, it is worth mentioning that academics start from a highly critical point of view, something that is fairly justifiable. Artificial Intelligence could be something extremely powerful. It could be even more powerful if it is given a high stand within society and a green card to act however it sees fit. In other words, to take the example of the enforcement of the law, by replacing human judges with Artificial Intelligence, AI now holds a strong position of power, capable of dictating the lives of humans. If this turns out to be the case, and as we shall see in Literature Review, it has already been implemented quite extensively, it is the job of the academics to conceptualize the problem, frame it, and pose it to the right people. In one sentence, Artificial Intelligence is for academia both a blessing and a curse. But what exactly is Artificial Intelligence?
**Artificial Intelligence Definition**

“The art of creating machines that perform functions that require intelligence when performed by people” (Kurzweil, 1990, p. 14) is one of many definitions, but one I think encompasses all four aspects of Artificial Intelligence, namely Thinking Humanly, Thinking Rationally, Acting Humanly, and Acting Rationally as proposed in the book “Artificial Intelligence: a Modern Approach” (Russell & Norvig, 2010, pp. pp. 1-2). In other words, Artificial Intelligence is any computational method -even though algorithms are solely addressed so far- that can give a human-like approach to a machine. Even though the title incorporates Intelligence, the term does not argue specifically that it reproduces a human brain, but rather, through the use of Logic and mathematics, it achieves for a machine to not be solely an operator of commands but create some sequences of its own. To put it even more plainly, Artificial Intelligence does not refer to the machine carrying the operation (irrespective of the machine’s resemblance to the human form) but rather to the algorithms hidden in its core, powering its every function. It is, in this sense, perfectly rational to claim that a software that has no machine parts whatsoever is indeed Artificial Intelligence.

On this confusion, James Barrat in his book “Our Final Invention: Artificial Intelligence and the End of the Human Era” claimed that at this moment researchers agree that three different stages of Artificial Intelligence exist or might exist in the future (Barrat, 2013): Artificial Narrow Intelligence (ANI), Artificial General Intelligence (AGI), and Artificial Super Intelligence (ASI). A brief description of each is due, in order to set a clear ground for my arguments later in the thesis.

*Artificial Narrow Intelligence* refers to any Artificial Intelligence that is exceptionally good at a certain task but lack the ability to perform in any other context (Barrat, 2013). This kind of AI is very much incorporated into our world already, and one need not search for them, for chances are that she is interacting with them on a daily basis. To give a few quick examples, Google Maps, an Artificial Intelligence algorithm by Google is probably a form of AI that most citizens in western countries have experienced. This algorithm is excellent at searching for the best routes, even predicting traffic or a cancelled train, but it will not be able to cook a meal, do one’s shopping, or play competitive chess. Deep Blue, on the other hand, is another very famous Artificial Intelligence algorithm that have occupied the press, when it managed to defeat Gary Kasparov—a chess grand Master- in his own game.

*Artificial General Intelligence*, on the other hand, refers to any Artificial Intelligence algorithm that is as competent as a human being, put in any position that a human would be. In other words, AGI is the term used for the very notion of humanoids; machines -or robots- that will act as humans and will interact with other humans on an equal level. This level of Artificial Intelligence has not yet been achieved, but the interest of researchers has been turned towards this direction. It is important to state here that ANI is considered of rather basic level of difficulty compared with AGI. The leap is simplified in Barrat’s three stages, but in reality, achieving AGI is mainly out of reach at this stage.
Artificial Super Intelligence refers to the hypothetical scenario that AGI is indeed achieved, and manages to create its own version of itself, accelerating in technological prowess, to a point where it will surpass human intelligence. There is not an example I can provide here, as this notion should be beyond our understanding. If I may use a crude analogy, while a monkey can be trained to use the internet, it is impossible for it to understand the concept of the internet. In the same sense, even though humans might see and interact with ASI, it will be impossible to a) predict what it will be capable of doing, b) understand what is doing in real time. This is, of course, merely speculative at the moment, but a lot of academics try to imagine if this will be humanity’s doom or salvation. Nick Bostrom’s “Superintelligence: Paths, Dangers, Strategies” is an excellent read that analyses similar scenarios, as well as the actions we might have to take a lot earlier than that, if we want to avoid gambling the fate of humanity to the will of the Artificial Super Intelligence (Bostrom, 2014).

The main focus of the thesis will be the first category, Artificial Narrow Intelligence, for two main reasons. The first one is that, in my opinion, the technology that our society focuses at the moment should be under further scrutiny, because if the foundations are not laid while still in the beginning, it might be even more difficult to do so as technology progresses. The second reason is that, for a thesis on Artificial Intelligence Ethics, a clear notion of what the arguments are about is essential. As the reader will discover shortly, ethics is a very ambiguous field in itself; trying to incorporate it into a subject that has not yet seen the light of day might prove an impossible task. Artificial General Intelligence and Artificial Super Intelligence are the subject for further research, if these ideas ever come to fruition. For the remainder of this project, whenever Artificial Intelligence is used as a term, it will refer only to Artificial Narrow Intelligence.

Ethics and AI

Let us move now, to a philosophical branch that was part of the formal discipline of philosophy for many centuries, but which has found a renaissance of sorts inside the Artificial Intelligence regime; ethics. Different ethical theories, their histories, doctrines, applications and criticisms will be discussed in the next session Literature Review where they will also be put under different schools to make my arguments more coherent. What I want to address now is why are ethics such a hot-blooded debate among recent scholars, and what are the main themes that are easily identifiable. The existence of such a debate might not be shocking since AI can potentially be humanity’s greatest invention. However, strikingly enough, most

---

3 One might argue that a thought experiment could be a way to tackle this problem. However, as mentioned before, since this essay tries to go beyond philosophy’s practice for arguing’s sake, real, concrete examples are hard to come by because nothing similar exists in the real world. Therefore, creating an accurate simile would be impossible.
proposed models so far have been either near-sighted or too normative to address the question effectively. My model aims to navigate the discussion in virgin lands, that of relativism. For the purposes of this paper, the whole aspect of ethics in Artificial Intelligence has been condensed into three main categories; three categories that are not worlds apart, but most -if not all- of the academic literature falls under, one way or another. It will be apparent that the three categories that I will address next are not simply philosophically abstract concepts but real, concrete hypotheses that have direct ethical implications.

The first principle concerns the right of the machines. What does make a human being a human being is something that has preoccupied philosophical research for an extensive period of time, and the results are not satisfactory enough, at least if one wants to have an objective definition. Ideas range from the Pre-Socratic notion of fire within, to the Christian divinity, to the Descartian soul, etc. The struggle, then, for agreeing upon a machine’s rights privileges is completely understandable. In a way, scholars try to tackle a novel problem without firm ground to stand upon in order to make their claims (Coeckerbergh, 2010).

Some researchers think that machines should indeed have moral rights if they achieve Artificial General Intelligence, because this is the bottom line of the ongoing debate about the originality of humans (Coeckerbergh, 2010; McNally & Inayatullah, 1988). To put it in simpler terms, some philosophers claim that what distinguishes from other species is our ability to think logically, irrespectively of the title they assign to it, a notion they call “Anthropos”. Therefore, if machines achieve a similar amount of intelligence, we ought to hold them as equals in our society, even though they might not even have a physical form.

Others discredit this idea, as they think that Artificial Intelligence, even if it is Intelligence, it will always be an offspring of humans. As a consequence, they should be held at a lesser level, because the discussion of what constitutes a human being has not yet been resolved; putting every claim under the same umbrella of abstract logical thinking is too simplistic (Gunkel, 2017). At the same time, humans can’t be certain that the survivability of the species will be the top priority of the AI agents, even if the opposite is true. That is, while we may go into great lengths to ensure that algorithms will be held to the same standards as humans, said algorithms might not think the same for humans and undermine us, even if unwillingly.

The second principle that emerges right off the bat, one that has been briefly touched upon above, discusses the power relations between the inventors of the Artificial Intelligence agents and the rest of citizens. It is rather apparent that these agents are combining utility with an immense amount of power in the hands of their creators (Boyd & Holton, 2017). One should only imagine that Big Tech -if these companies are indeed those that will introduce us more consumer-friendly Artificial Intelligence- will be able to intrude AI into every house,

---

4 Again, this is related to posthumanism, but unfortunately there is not enough space to argue for the transition.
every company, every public building. At the same time, these Big Tech companies will have an absolute control over how these agents interact with the rest of the world (Boyd & Holton, 2017). It would not be far-fetched to assume that they might use this opportunity to further establish their superiority and promote their hidden agendas.

To give one example with Google’s application Google Maps, and its vast usage and utility in the western world. One need only input their final destination and enable their location services and the algorithm will give them the best route if they drive based on their preferences (shorter distance, no tolls, no traffic, etc.), or the best alternative option to take if they use public transportation. Let us hypothetically assume now, that Uber wants to promote their business, but do not want to do it through regular advertising but has an alternative proposal to Google. Rather, since both companies are Silicon Valley based, they decide to promote Uber through Google Maps without explicitly declaring so. How difficult would it be for Google to tweak its algorithm a little bit, just enough to make the best alternative to public transport an Uber ride?

This is just an example, and I do not want to imply that Google, Uber, or any other company is doing under-the-table deals to promote their services or products to get an unfair advantage. What I want to state, though, is that this is something that could potentially happen, therefore caution and possibly proactive counter measures are encouraged. It is -I hope- now clear how power relations can have real ethical implications both in our public and private life, as well as our everyday encounter with these technologies.

The third and final principle concerns the morality of said Artificial Intelligence agents themselves. Many critics point out that morality is something that has to be inserted into these machines, lest society wants to abide by the rules of heartless algorithms. Several scholars have pointed out that AI agents are already incorporated into our society, sometimes holding positions of authority. Judging criminal cases and evaluating the potentialities of schoolteachers are just two examples that Cathy O’ Neil brings to light in her book “Weapons of Math Destruction” (Neil, 2016). These algorithms may be advertised as objective and factual, but some things that they are surely missing are compassion, humanity, and morality. One can be certain of that because they are promoted as agents that will take away the subjectivity -and therefore fallacy- of human judges (in both cases) and restrict themselves to the actual facts at hand.

The problem is that in most cases human morality and ethics are essential, or at least they have been up until now. These algorithms, if they are to hold such important roles and be entrusted with prolonging democracy, have to not only rely on mathematical models but encompass some kind of ethical guidelines. Some critics, especially coming from entrepreneurs and technological determinists claim that the utility of those agents is that they have no morality, because this is simply the point; why putting an Artificial Intelligent machine to replace a human if the means are going to be the same? The result will probably
be the case as well, rendering these machines more human-like but insufficient for the task at hand.

Further elaboration will come on the section *Literature Review*, as the latter is an argument that has been widely discussed among the academic cycles. This is the single most debatable and extensively researched topic on Artificial Intelligence Ethics, and the one that this project will try to tackle. But before I move on to the discussion of the literature, I would like to note some actions that governments and companies have taken to address the last two principles. Big tech companies such as Facebook Inc. has launched its own ethics board committee, while the state of California in the USA has not only abandoned but banned facial recognition cameras and its use by the police.

**Introductory comments**

These might seem very noble acts and well-timed decisions, but Rodrigo Ochigame has a different view on the matter. In her article “The Invention of “Ethical AI”: How Big Tech Manipulates academia to Avoid Regulation” (2019) in *the Intercept*, she claims that even though the façade of projects such as Facebook’s “ethical AI” is one of transparency and truth searching, in reality lobbyists from Big Tech are the one driving the wheel (Rodrigo Ochigame, 2019). She gives the example of Eric Schmidt, which holds a position inside the committee while at the same time was -during that time- Google’s CEO. Lobbying poisons projects such as these claims Ochigame, who has been working with Joichi Ito, which was involved in a quite big AI scandal, for an extensive period of time.

The issue gets deeper, as Ochigame writes that academia is deeply connected into this scheme as well, with certain professors being labeled as “experts” on Artificial Intelligence matters without having formal education on this discipline. For her, these professors exist to solely promote Big Tech’s benefits. The hardest piece of evidence, she claims, is the fact that these Non-Governmental incentives are mostly sponsored and promoted by Big Tech’s budgets. The reason for that is simple. There are three possible scenarios if and when people ask for regulations on Artificial Intelligence:

A) there will be no change. The movement for change will not be strong enough, and Big Tech will grow so big that it will be impossible to regulate. B) there will be slight change. The movement might be strong enough to cause some ripples but not enough to shake the industry in its core. The result will be a shallow, mild regulation that will only put the real problems under the carpet. C) there will be significant change. The movement will be strong enough to push for harsh regulation, democratization of research on Artificial Intelligence (open AI), and an end to the Big tech era. Ochigame concludes that the outcome for Big Tech are stated from best to worst. In order to prevent a possible C scenario, they censored themselves, giving an air of regulation more like B, while at the same time lobbying for A.
At the same time, the Department of Defense in the USA is also involved and is equally disinterested in having openness and fairness in Artificial Intelligence. This happens because a lot of military technology focuses on AI Research and Development. Algorithms that calculate the best routes for soldiers, unmanned drones, rockets that “neutralize” the most threatening targets, and many other weapons and tools are not meant to be shared. In a way, the DoD is actively pushing in unfairness, because warfare is an unfair struggle anyway; if they want to have the upper hand in modern warfare, they need to keep their research a secret. In the end, Ochigame claims, we have Non-Governmental Incentives that are backed up by corporate money, its members are promoting other agendas, and has to fight its way through the government as well.

The banning of facial recognition cameras in California, on the other hand, is something that is seen as a step forward. After reports came in of misuse of the data collected, of privacy issues and overall utility of the endeavor, the government decided that the risk of getting strong resistance and opposition was weighing way more than the benefits that the cameras offered. This is an example of the third scenario (C) in Ochigame’s model, one that pushes for real change and useful regulations.

I hope that these examples show the level of diversity that exist in the usage, the application, the utility, and the mentality of Artificial Intelligence agents, as well as some real ethical implications that might occur, given our impassivity in such an important matter. Let me now move on and begin my literature review of the main Artificial Intelligence ethical schools, their arguments, their criticisms, solutions, debates, and blind spots.

**Literature Review**

The literature review is split into three subsections based on three major ethical school of thoughts; deontology, consequentialism, and Aristotelian virtue ethics. While social scientists have provided a huge variety of arguments and different approaches, most of them can fit under one or several schools’ ideologies. This means that a certain scholar might appear in more than one section, but this is reasonable because the focus is on presenting the different arguments falling under their respective ethical school of thought. It might be possible, therefore, for a scholar’s complete work to cover several arguments falling in different categories. I will present scholars’ ideas and arguments based on one school’s ideology first, then provide some scholarly critique, and after move to the nest school. When all three schools are sufficiently explained and critiqued, I will move to the next section where I will relate all of them to ancient Greek philosophy.

Before I begin though, a word of caution is due; I am not going to extensively cover the work of computer scientists, or technical aspects of the creation of algorithms, because this goes well beyond the scope of this project. Rather, my literature concerns arguments made
from social scientists or engineers that have addressed more theoretical problems than modelling the agents.

**Deontology**

The first major school of thought concerning Artificial Intelligence ethics is Deontology. Its most prominent philosopher is non-other than Immanuel Kant with his famous “Categorical Imperatives”. In it, he had two main theses: a) ascribe moral values to specific actions (good, bad, etc), and b) that there are certain values that are absolute. In other words, independently of the case or surrounding context, one shall not lie or shall not kill (to give two prominent examples) (Kant, 1882). This notion of absolutism has influenced some modern scholars to argue about a potential deontological model in applied Ethics in Artificial Intelligence. In the following pages, I present five different main arguments that have Deontology either in their shaping or explicitly in their ideology. The reader should keep in mind Deontology’s tendency for absolute values as she reads through these five arguments in order to comprehend the common trends that they point to.

1. **Transparency**

Joanna Bryson and Alan Winfield in their article “Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems” claimed that ethics in AI are not only greatly needed, they should be implemented as soon as possible (Bryson & Winfield, 2017). To address the lack of firm ground for ethical frameworks, they proposed that transparency is probably the main ethical point that developers should work upon. Transparency can come in many different forms, and they discussed five: users, safety certification agents, accident investigators, lawyers or expert witnesses, and wider society (Bryson & Winfield, 2017). While transparency will be used differently by each category, there is a deontological approach as to the firm way of addressing the problem. In other words, transparency is seen as the absolute value, while any lack of it is seen as unethical.

Peters et al. discuss another implementation of deontological principles into AI agents, specifically into the context of internet-delivered therapy in digital mental health (Peters, Vold, Robinson, & Calvo, 2020). In this perspective, deontology is not only advised but actually implemented by the IEEE(Institute of Electrical and Electronics Engineers), where there is a double normative framework, responsible to incorporate the totality of potential uses and issues: a) the responsible design, where the developers will not only be enforced to create ethical AI agents but they will also be willing to do it, if they believe in the IEEE’s value system, and b) the spheres of technology experience, where users will be able to select the level of access they want to give to the algorithms (Peters, Vold, Robinson, & Calvo, 2020). Having a choice to trade between accuracy and privacy, users might choose a different set from an already pre-determined normative set of actions.
2. Artificial Intelligence omnipotence

Another significant argument that is put forward by quite a few scholars is that Artificial Intelligence models cannot and should not be used as omnipotent agents. As the level of abstraction for the creation of these models goes up, a trend is visible. A trend that reveals our human need for solid, concrete foundations. As people move away from religion, a new kind of omnipotence arrives. An agent that can judge a criminal case objectively, an agent that can decide for one’s loan unbiased, an agent that can evaluate teachers sophisticatedly. Interestingly enough though, even though Artificial Intelligence is still in its primal version today and clearly not even close to omnipotent levels, it is being trusted much more than a human to solve problems.

This, however, creates issues. While one can try to understand and appeal a case when a human judge is involved, it is impossible to do that with AI agents; this happens because they are -in essence- black boxes, machines that work on mathematical models that are not revealed to humans. As a consequence, one cannot know why this decision has been made, and therefore makes it impossible to appeal. As a matter of fact, errors can be produced from different angles (O’Neil, 2016).

Statistical Insignificance refers to the amount of data gathering for the function of the algorithm (O’Neil, 2016). If not enough data are collected, generalizations are not only impossible but dangerous if carried out anyway. Another angle that errors can occur quite easily is the fact that an algorithm is as good as the feedback it gets so it can tweak its calculations accordingly. But are errors without consequences bad? From a deontological perspective they are because an error is still something that should not happen. In many cases, however, there is no feedback given to the agent, therefore being prone of producing the same errors again and again (O’Neil, 2016). Yet another problem arises as some models are very easily inflated or gamed, even unintentionally, so that a question of robustness occurs (O’Neil, 2016).

O’Neil’s last argument about robustness has been extensively discussed in other sources as well. There seems to be a lack of trustworthiness in Artificial Intelligence agents for many different scenarios that AI is being used at the moment. Some of them include driverless cars (Marcus & Davis, 2019; Broussard, 2018; Fry, 2018), justice (Fry, 2018), medicine (Fry, 2018), crime (Fry, 2018), and evaluation of workers (Broussard, 2018; O’Neil, 2016). The first thing that comes to mind here is questioning the authorities for their ambiguous decision of using AI in important positions of power when it is clearly not ready yet. But decision-makers often do not see the whole picture; people are being used as cannon fodder for the training of AI because they are expendable or not valuable enough, just plain numbers for the authorities.
(O’Neil, 2016). If I can correlate it with Joseph Stalin’s famous quote “When one man dies it’s a tragedy. When thousands die it’s statistics”, I shall⁵ (Antonov-Ovseyenko, 1981).

3. Artificial Intelligence Abstraction

One more problem with Artificial Intelligence agents is their level of abstraction. In other words, in order to tackle more real-life situations, developers of those agents have to make algorithms that are much more sophisticated, much more complex, much more abstract. As a consequence, as the level of complexity rises, so do the possibilities for unaccounted mistakes (Kerns & Roth, 2019). This notion is not very difficult to grasp, as this is purely basic mathematics. For the sake of argument, let’s hypothesize that for each line of code, the chance of a bug or a logical mistake is X. If our agent has only one line of code, we can expect them to be 100% accurate - X. The more lines of code, the number of mistakes rises. For twenty lines of code the expected outcome would be 100% - 20X. In order to build sophisticated algorithms, researchers write tens of thousands of lines of code. It is easy, therefore, to see that the more abstract the model, the easier it is for mistakes to occur.

But the bigger problem is not the occurrence of problems, but the position, discoverability, and abstraction of them. In other words, researchers might expect unexpected problems to occur, but what they cannot anticipate is where these problems will arise, and if they can be spotted at all. O’Neil gave an example with an algorithm deciding people’s mortgages. While this model has been tested thousands of times before entering the market and proving flawless, a trend was observed right from the start; it discriminated against certain tax numbers. (O’Neil, 2016). This happened because of a small bug that was not tested or thought insignificant from its developers, which proves Kerns’ and Roth’s argument that the more abstract an algorithm is, the easier it is for unaccounted mistakes to occur.

But Kerns and Roth went a bit further than that. They claimed that scientific discoveries focused solely on data, fall into a pitfall; that of not understanding real-life’s complexity (Kerns & Roth, 2019). They go on to provide several examples of discoveries in medicine, physics, and chemistry where all of them had one similarity. Scientists gathered large amount of data and interpreted them in a way that would justify them. To put it differently, they identified patterns in the collected data, and they hypothesized the results. The problem is that these results were never tested under different circumstances and this created faulty information.

The same thing happens with Artificial Intelligence agents. They get trained into large datasets (a technique called Machine Learning) so that they identify patterns invisible to humans, therefore being able to perform a given task much better than said humans. But

⁵ This quote has been questioned as an original Stalin quote, but my point here is not to argue for historical accuracy. The saying is more important than the person attributed to.
Kerns and Roth do not share this optimism. For them, irrespectively of the amount of data a researcher feeds to an Artificial Intelligence agent, having no way to test what the emerged pattern is or training the algorithm in more than just datasets -like real life simulations f.e.- underestimate the complexity of real life (Kerns & Roth, 2019). Having abstract machines trained solely on datasets will result in circumcised machines; machines that will have minimum application in contemporary societies.

4. Artificial “Intelligence”

So far, the arguments presented have taken as a given the assumption that Artificial Intelligence agents are indeed intelligent. While they seem like they are thinking like humans do, this is only part of the bigger picture. Broussard starts from building the simplest of programs, “Hello world”, in order to argue that machines don’t think like humans; they simply operate mathematical calculations (Broussard, 2018). “Hello World” is something of a ritual for every coder. It is the simplest program one can write, and it exists as a single line of code. But even from this tiny algorithm, Broussard concludes that machines operate on a completely different level than humans. Instead of going through a cognitive process and produce an outcome that will be conscious and willing, algorithms just replicate what they have been taught. While the outcome is the same, there are two main differences: will and imagination. Humans must want to say or write “Hello World” in order to do it and they need to think of a way to do it. Machines do none of the above.

This distinction is important for Broussard because people can distinguish between ambiguous words like good and popular, but a machine cannot. She gives the example of racism and speed limits to prove her points. Racism is popular - at least among certain communities - but it is generally accepted as not good. At the same time, speed limits are good but are not popular (Broussard, 2018). The fact that I do not need to further elaborate on these examples shows that humans can understand the difference in ambiguous words; Artificial Intelligence algorithms cannot. This is extremely important if the widespread application of AI agents is taken into account. Artificial Intelligence developers consume a larger amount of energy trying to make the public believe that their agents are smart rather than trying to actually make them so (Marcus & Davis, 2019).

For Marcus and Davis, AI agents are still quite narrow. They can only specialize in one task; by trying to stretch their contemporary capabilities further without development is a mistake. A possible reason why these algorithms are still quite narrow is that they cannot comprehend natural language (Marcus & Davis, 2019). Surely enough they seem like they do, as Alexa, Siri, and other commercial AI agents try very hard to prove, but they do not really understand natural languages; rather, they understand certain keywords, phrases, or patterns of speech, just enough to make easy, mundane tasks automatable (Marcus & Davis, 2019). But that is as far as they can go. To claim that AI agents can comprehend natural language the way a human does is a bold statement, to say the least. It is, therefore, quite apparent for Marcus and Davis that machines do not think; they make us believe that they do.
Another argument coming from O’Neil, proving that algorithms do not think, is the fact that a successful algorithm gets trained on a large dataset, then it is tweaked to be optimal, then is being given another dataset, and the process repeats. The fact that we should constantly tweak the algorithms in order to fix them proves that they do not have autonomy (O’Neil, 2016). To supplement her argument, Marcus and Davis claim that algorithms are tweaked only after the results are wrong (Marcus & Davis, 2019). In other words, developers do not expect the algorithm to adjust itself optimally, so they need to constantly correct them. This proves that AI agents have no critical thinking, something essential in natural language comprehension.

Broussard takes as a given the previous claim that machines do not think and enriches the discussion by claiming that Artificial Intelligence agents create big problems in many situations, because they are not “intelligent” (Broussard, 2018). She gives the example of autonomous cars and the way they behave, which is nothing more than following a list of instructions. Useful as it may be as an example of Game theory -if all cars abide by the law, the unpredicted situations would be far fewer- it fails on two levels: a) in these unpredictable situations no one knows how the machine will react, b) it is yet another proof that machines are simply following commands.

In addition, social biases occur when feeding human-generated data to the machines (Marcus & Davis, 2019), since these algorithms are then going to be used by banks, schools, universities, the government, companies, etc. Broussard writes that biases in terms of racism, sexism, elitisms, and more, are taught to machines because the algorithms have been trained on human-collected data. In other words, Artificial Intelligence agents would be much fairer, had they been trained with data collected from a perfect society. Jane Wakefield reported on BBC on 2016 of a Microsoft bot, called “Tay” that was supposed to learn from conversations and interactions in Twitter, and create content mainly for young adults. It took only 16 hours for Tay bot to be officially shut down by Microsoft, after it had committed major offences, supporting hate speech and genocides among them (Wakefield, 2016).

5. Ethical Framework Deontological model

Luciano Floridi, a prominent scholar in AI ethics and a professor at Oxford University, claimed in his article “AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations” that a deontological model can provide humanity with four opportunities (if taken as such) or threats (if not being materialized adequately): a) who we can become, what we can do, what we can achieve, and how we can interact with each other and the world (2018).

Building up on this argument, he ended up with four main themes that AI ethicists must try to address: beneficence, non-maleficence, autonomy, justice, and (a last one which he added later) explicability (Floridi, et al., 2018). Interestingly enough, all five themes are based on a deontological perspective, as it must be clear what autonomy or justice is for
example, if ethicists hope to find common ground. The utility and influence of deontology in this model cannot be understated for an additional reason though; the basic opportunities or threats, as stated above, are based on such absolute values themselves. As a consequence, starting from deontological foundations will result in deontological solutions.

6. Diversity is scarce

Another issue that Garcia-Cuesta and O’Neil bring to light has to do with the fact that by using Artificial Intelligence agents we eliminate diversity that would exist in the workforce (Garcia-Cuesta, 2020; O’Neil, 2016). Most often than not, recruitment practices have changed over the last years. There is an essential first step in contemporary practices, where algorithms scan the collected CV’s and sort them out in terms of applicability to a certain position. It is only after this stage that human recruiters might come into play; in some instances, the whole process is completely automated.

This is especially worrisome because diversity in employers is evaporated. While a CV might be considered inappropriate for one recruiter in the past, it might as well be in complete accordance with the next. Therefore, different people having different CVs were not only a bad thing, but highly encouraged socially, as this was a double elimination from both parties because common attraction had to be established in order to pass to the next stage. This process was making sure that potential workers would eventually find an employer that could work under, irrespectively of the uniqueness of their character and qualifications. This diversity is not easy to maintain in contemporary societies, because the algorithms scanning the CVs do not have opinions, subjective criteria, or things they like. Rather, they operate in a completely isolated, mathematical way, making sure that the calculations are correct, and the people with the CVs most fitted to the algorithms may proceed to the next round. It is important here to highlight that the CVs that are chosen are those that are most compatible with the algorithm, not the employer or the job at hand.

Another issue with the loss of diversity is the fact that Artificial Intelligence agents do not allow for flexibility and understanding; their will is absolute (O’Neil, 2016). In other words, some people might have certain minor disabilities or any other difficulties in interacting with the algorithms. Instead of humans empathizing and possibly try a different method to communicate with those people, Artificial Intelligence agents do not succumb to these needs. They have a single way of doing things, straightforward. If a person from a minority group fails to interact with them, it is either their fault or a statistical nuisance that should not be taken into account. These arguments touch upon several other themes that I have covered so far, like the issue of agents using people as cannon fodder for proving their utility and those agents’ inability to think critically or approach the subject at hand from a different perspective.
7. Discrimination benefits

Advertisers faced big problems with audience targeting. Before the introduction of the commercial internet, it was rather hard to get detailed analyses of a medium’s audience demographics or similar information. The introduction of the internet solved this problem for marketers, as now they both had very detailed reports on their consumers and a medium to target them directly without any spillover, as it happened with traditional media such as radio and television. The previously discussed discriminatory algorithms, therefore, not only are they not tweaked to combat segregation, but this discrimination is instead desirable. The existence of biases in Artificial Intelligence agents is known to Big Tech, but not only did they not try to fix them or faced repercussions, but created marketing opportunities out of them (Benjamin, 2019).

A very interesting example is Google’s algorithm for their search engine, called PageRank. This algorithm prioritizes what is most shared or created the biggest buzz, something fairly convenient for an ad-based company (Noble, 2018). Google never intended to cross-check what is being shown on top of their search results. Validity scored lower than accuracy, and while this may be convenient for some users, it creates big inequalities for producers. In other words, this algorithm is created to deliver viral content based on the keywords. It is not hard to imagine how this discriminates in favor of those screaming loudest. In most cases, as Noble argues, the search results end up having discriminatory content because this is the tradeoff for popularity; marginalization of the weak (Noble, 2018). As certain races are assumed to have more spending money, more content is created for them, therefore content created for other races gets more and more pushed aside by the algorithms themselves6.

But there is another reason why the discrimination biases hurt democracy; for the first time in history in official democratic societies, there is corporate control over public information (Noble, 2018). Big Tech does not control the documents or the archives themselves, but they do control the medium through which most citizens access these public documents. This is especially important for information that is of vital importance, like official medicinal reports, official governmental statements, or scientific essays publicly available. Instead, in many cases the algorithm will prioritize blogs or commercial sites because they have found a way to game the algorithm either by SEO (Search Engine Optimization), or being referred by other commercial sites, or even paying money to appear on top. I think it is too much to comment further on this issue. Suffice to condense everything in a single sentence;

6 I am aware that in Sweden races are not a big part of the academic discussion. However, since (Noble, 2018) is American and focuses extensively on the American context, the racial issue is very relevant indeed.
commercialization, not only favors from discrimination, it benefits from marginalizing important content over trivial.

**Critique of Deontological Ideas in AI**

Even though certain scholars have been positively influenced to introduce their own solutions to the problem by the school of thought of Deontology, not everyone is equally convinced of the latter’s application to this problem. Korsgaard claimed that such a model is deemed to fail because it fails to identify the actor behind a person’s actions (Korsgaard, 1999). To put it differently, regulating the industry by compelling developers to incorporate absolute moral laws into their algorithms, a dual problem occurs: a) problems might occur when said moral law contradicts basic human rights, because this is something that might happen in real life, and b) in such a case responsibility is very hard to put upon anyone. Developers have abided by the law, and so did their Artificial Intelligence agents.

A second scholar that found deontological studies problematic for AI is Milan Tasić. He argued that objectivity incorporates logic (Tasić, 2015). In other words, to render something objective would require logic, while render something subjective is impossible with mathematical thinking. All in all, mathematics is the language of facts. While this might seem desirable in Artificial Intelligence agents, especially since, as I explained before, they operate through Logic, reality is more complex. Making objective ethics is like playing with fire; it is not an accident that dozens of different ethical theories still exist in 2020. It would be impossible to choose one morality to apply for all algorithms, even if this is the easier solution. Even though deontology seems like a good fit to the puzzle, it is very conformistic and narrow in its approach for such a wide phenomenon as is Artificial Intelligence.

Marcus and Davis are not as optimistic as other scholars I cited so far when it comes to the transition from Artificial Narrow Intelligence to Artificial General Intelligence. AI requires a tremendous leap; trying to speed it up can prove disastrous (Marcus & Davis, 2019). While humans are still indecisive if they should trust ANI for anything more than a tool, the introduction of AGI will obscure this discussion on multiple levels. For starters, it will not be clear if machines are going to replace humans on intellectual jobs, and not just the ones that are easy to automate, as is mostly the case at the moment. Then, an issue might arise with a certain AI agent holding multiple positions, since this is what AGI originally promises. After, it will be unclear if humans should trust these machines, because by achieving AGI it means that they will be on a similar intellectual level with humans, therefore they might wish to promote their own agendas. Further complications might arise when a decision has to be made about these algorithms creating their own algorithms.

In a nutshell, when Artificial Intelligence becomes truly intelligent the whole discourse has to be reshaped. But social scientists should rest assured that this will not happen anytime soon (Marcus & Davis, 2019). This level of intelligence proves to be completely out of reach at the moment, to the point that developers cannot even predict if it will ever be achieved.
This is where Marcus and Davis ring the alarm bells. If developers become greedy and try to oversell something which is clearly out of hand at the moment, the consequences will be dire. Trusting machines that work is a debatable subject on its own; trusting machines that are not up to the task should be impermissible. The damage done to democratic societies, the level of citizens allowance and trust on Artificial Intelligence novel projects, and the human cost are just a few examples that are most vulnerable by making hasty moves and not standing onto firm foundations.

**Consequentialism**

The second major school of thought concerning Artificial Intelligence Ethics is called **Consequentialism**, mainly -but not only- represented by the ethical theory called **Utilitarianism**. Other forms of consequentialism include ethical egoism and altruism, motive, negative, and state consequentialism, as well as teleology (Scheffler, 1998). Utilitarianism has been split into two categories: act utilitarianism and rule utilitarianism, and their representative proponents disagree on quite a few arguments. To keep this discussion within the space limits, however, as well as keep the project concise and to the point, I will only refer to utilitarianism in its main concept; the promotion of maximum happiness and pleasure in the affected participants (Scarre, 1996). The reader thus, should keep in mind consequentialism’s main argument which is based on the simple notion of judging the morality of an action by its outcome.

1. **Different consequentialist approaches**

Amitai Etzioni and Oren Etzioni in their article “Incorporating Ethics into Artificial Intelligence” compared between two main ethical fields, a top-down and a bottom-up approach. The top-down approach concerns decisions implemented by the developers themselves to the algorithms, while a bottom-up approach concerns the agents learning from observing humans (Etzioni & Etzioni, 2017). The second approach, they argued, is not optimal, as it will teach agents what is common rather than what is ethical. This is a classic consequentialist approach, where the maximum amount of happiness cannot be left to sort itself out in society. As societies do not usually act in a consequentialist way, by allowing algorithms to learn from everyday patterns will work against an ethical theory that has to be implemented by its designers. To put it plainly, if a utilitarian approach is to be implemented, it has to be via a top-down system, where the algorithms will explicitly incorporate happiness maximization.

Another hint that reveals Etzioni’s and Etzioni’s lean toward consequentialism is the fact that they argue quite extensively about the “trolley problem”. The trolley problem’s main form is thus: five people are tied onto some railway tracks and a train is fast approaching. A person has a lever available, able to alter the course of the train to a second track that a single person is similarly tied. The dilemma is if said person should alter the train’s course and kill one person instead of killing five (Foot, 1967). In consequentialist theory, the answer is
obvious; the lever must be pulled, and the single person sacrificed for the maximum amount of happiness will, this way, be achieved. Because, however, this outcome is quite debatable among social scientists, Etzioni and Etzioni argued that it is not a valid criticism to consequentialism because these occurrences are quite rare. Giving the example of autonomous cars, they claimed that in the same situation, if not an algorithm, a human driver would be on the spotlight. Since there is no objective answer as to what ought to be done by humans, it would be irrational to judge a machine based on an ambiguous thought experiment.

Jarrett Zigon, based on Levinas’ work (2011), claimed in his article “Can Machines Be Ethical? On the necessity of Relational Ethics and Empathetic Attunement for Data-Centric Technologies” that what is good and what is moral, when concerning Artificial Intelligence, is the idea of the other (Zigon, 2019). In his words, “the Other always exceeds any idea of the Other that ‘I’ may possess—for example, in terms of what and who the Other is, what the Other desires, or even how ‘I’ should act with and toward the Other” (Zigon, 2019). This is a central consequentialist idea, called relational ethics. An oversimplification of it would be thus: if all agents thought of how to minimize harm for other agents, then all would benefit. It is, in a way, the exact opposite of egoism, flirting with the Aristotelian notion of altruism. In this context, by taking into account the social issues discussed in the Literature Review, agents will have as a goal to, for example, promote maximum diversity. The concept of a normative path towards a specific goal is clear, which is the reason why relational ethics fall under consequentialism.

Derek Leben argued in his book “Ethics for Robots” that “such algorithms should be evaluated by how effectively they accomplish the problem of cooperation among self-interested organisms, and therefore, rather than simulating the psychological systems that have evolved to solve this problem, engineers should be tackling the problem itself, taking relevant lessons from our moral psychology “ (Leben 2019 p. 2). Using John Rawl’s ethical framework, this consequentialist approach focuses on the “maximin principle”, which is the action that maximizes the minimum value. In other words, going against the utilitarianist theorem, Rawls and therefore Leben, argued that the choice of which action should be taken is the one that has the least negative impact to people who will be worst off it. Taking Artificial Intelligence in this context, Leben proposed that since marginalization is impossible to abolish through the use of such agents, the alternative of the lesser evil should be adopted (Leben, 2019), which is to have as little of it as possible.

2. Data Issues

Data mining has been researched on its own quite extensively. The arguments presented here will not cover the topic in general, but simply the overlap between data mining and Artificial Intelligence consequentialist ethics. While it seems that most issues that have been researched by scholars on data mining can be linked to AI research, some arguments follow the opposite direction; stemming out from Artificial Intelligence research but concern much
wider issues than simply the agents and the algorithms. It is the latter category that this section is about.

Fry argued that data mining is problematic for various reasons (Fry, 2018). One of them is the fact that companies will make certain to gain even the tiniest piece of information about their consumers. This constant squeeze of data is as enhanced by Artificial Intelligence agents as it is necessary for their survival. In a way, these data are the lifeline of said agents, because they need vast amounts of datasets in order to train and get “smarter”. Therefore, it is of vital importance for companies to extract even worthless pieces from users in order to support their algorithms that will then extract more data; a perfect vicious cycle (Fry, 2018).

This notion has found a commercial application as well. The company “Palantir”, while not very popular or known within consumers, is arguably one of the most famous companies in Silicon Valley. It markets itself as a “data broker”. In other words, what Palantir does is extract data from users through various means (apps, websites, cookies, etc.), creates datasets, and then selling those datasets to interested parties; mainly AI companies because they need them to train their algorithms (Fry, 2018). This might not seem troubling for some. As Fry explained through the example of ratings of products, consumers and citizens use data themselves. If one wants to purchase a product, chances are that she will advise the comments in a website or the star-rating system that has come to life from collection of willingly-given data for common use (Fry, 2018).

Issues, however, arise when companies having much more sophisticated algorithms that can analyze the data, can go much beyond the simple linear connection consumers do. For example, a human might think that since product X has 4.6 stars based on thousands of reviews it will be a better purchase than product Y which has 3.5 stars based on a similar amount of reviews. Algorithms, however, partly because they are made to incorporate a lot more variables, partly because they have much larger and richer datasets available, are able to calculate much more important conclusions. Based on a person’s address, mobile plan, usage of data at certain hours of day and location, and browsing history can predict a person’s race, economic status, job, affair, and many more. Then, other Artificial Intelligence agents use these results for micro manipulation (Fry, 2018).

O’Neil termed this phenomenon “the marketing of predators” (O’Neil, 2016). Big Tech companies have allowed for extensive profiling of users that could not only be used for gray-zone marketing purposes but for nefarious use as well. Targeted ads, for example, are one way for Big Tech to make money. Advertisers are happy because they can target only specific groups of people and minimize budget-spillover to uninterested consumers, while some users are happy because the ads are, for the first time relevant and interesting. While this is a big debate on its own, more important issues come to mind. Scandals like Cambridge Analytica (Confessore, 2018) have proved that these commercial practices can undermine democracy itself. In the case of Cambridge Analytica, a very small team was able to reach millions of
people with the help of Artificial Intelligence. It is apparent, then, that as algorithms become more sophisticated, profiling can become dangerous for various aspects of citizens’ lives.

The second main section with Artificial Intelligence and data regards how these data are being collected. As I have previously discussed, without a constant feed of large datasets, Artificial Intelligence agents are useless (Marcus & Davis, 2019). However, even if data are collected and then used as training grounds for algorithms to improve, several implications occur.

The first issue, as Marcus and David write, comes with the type of data that these agents are using. Using again the example of the star-rating system for the purchasing of products, it is quite apparent that in various situations the data have been either inflated or gamed. In other words, malicious actors have testified false information in order to promote their own agendas. These people (or bots) might be hired from the seller of the product to upvote it with multiple accounts in order to higher up their score, or from a competitor that wants the exact opposite. Irrespectively of the intention, the way and the result are the same; gamed data that are unreliable.

It has to be noted here that while humans possess critical thinking abilities, that cannot be said for Artificial Intelligence agents. Therefore, in cases where people intentionally alter their data in order to “confuse” the algorithms, are messing up with the core of the AI system; their ability to be trained properly. In the end, what most of those agents end up with is a mixture of real and artificial results that are impossible to separate. Developers should take numerous extra steps and jeopardize whole projects if they want to be accurate with the data that their algorithms are being trained upon.

This means delays, loss of deadlines, and possible monetary losses for their companies. So they hide this issue under the table and feed the algorithms the inflated data, hoping that it will only be a small percentage, insignificant enough in the outcome since the law of great numbers in mathematics suggests that the higher the quantity, the higher the chance for accuracy.

This is but one example proving Broussard’s point that in order to teach AI agents to work ideally, they should be trained upon ideal datasets. This would create other problems as these agents would not be ready for real life society, since human society is not perfect, but there is a big underlying notion here; humans can critically think, something which algorithms cannot do at the moment. It would be smarter to work toward that goal than trying to feed them data that are of poor quality and pray that the results are not going to be disastrous.

A final issue arises with O’Neil’s and Marcus’ and Davis’ argument that algorithms that get trained on data they created themselves create echo chambers. As I have previously mentioned, there is a vicious cycle of Artificial Intelligence agents that collect data from their daily interactions with humans. Then they use these data to tweak their algorithms and be optimal, repeating the process the next day. This is called an Echo chamber; a situation where
the same kinds of data are being repeated over and over again (Barberá, Jost, Nagler, Tucker, & Bonneau, 2015).

Interestingly enough, it has been observed that agents tweak slightly their algorithms the first day, only to do the exact opposite tweak the following day, ending up in the same position as two days ago (O’Neil, 2016). This has many consequences, from delaying the delivery of optimal agents, to agents that claim they are getting better every day without actually being so, to more and more databases being created uselessly.

3. Hidden from public view

Artificial intelligence development and operation takes place behind closed doors; as a consequence, it renders the discussion about its potentialities and applications quite challenging (O’Neil, 2016). For O’Neil, operating from behind the curtains is understandable for two main reasons. The first one is that had it not been so, people would know how these algorithms operate or would speculate with a higher success rate than they do now, and they would possibly try to game them. This could be with malicious intent like cyber-terrorism, hacking, data-breaching and many more, or with more benign goals in mind like, for example, a company ranking higher on Google searches, a producer’s video entering the “trending” list on the appropriate distribution platform or even enter the correct input so that one can guarantee a successful mortgage application.

The second one is that these algorithms are copyrighted or patented by their developing companies. In other words, it’s their “secret sauce” in the colloquial fast food language. It is what makes their product unique; opening it up for everyone to see would demolish their competitive advantage as well as their monopoly rights in their own creation. O’Neil, by providing these arguments, gives rise to the other side of the debate, the commercial one. In a way, companies would be discouraged to invest in new technologies if they knew that their findings would not be used solely for their advantage but would be open to the public for anyone to benefit.

While these arguments are solid and difficult to disagree with, other issues arise on similar grounds. By obscuring the details, Big Tech made it impossible to citizens and consumers to know how fair their Artificial Intelligence agents truly are (O’Neil, 2016). Even though the companies go to great lengths to promote their products’ fairness and honesty, without concrete evidence that these agents are working the way they are supposed to, the case is difficult to be made. At the same time, several incidents of algorithmic injustice relating their services harm them on a much bigger scale than a successful advertising campaign.

Also, locking the algorithms away from public view creates a second problem; without knowing how these Artificial Intelligence agents work, in the cases where they make mistakes—and such cases exist more often than companies would like to admit—without having knowledge on how they operate, it is impossible to contradict them (O’Neil, 2016). People
that have been negatively affected by algorithmic decisions are ended up facing a wall, because arguing that the results are wrong in any way requires some knowledge on how the algorithms came up with these results. Pleading their case with the agents themselves is obviously impossible as well. In the end, they are end up in a situation where they know they have been wronged but cannot find justice because no one knows how these algorithms came up with the conclusions. For the authorities involved, the agents might as well be right, even if the affected people do not want to admit it.

4. A Cheap Alternative

Under capitalism every service has to be paid. Ideally, the more expensive the service is, the more valuable it is in terms of the outcome the buyer receives. This is exactly the notion that Artificial Intelligence agents’ developers thrived upon; they offered a cheaper alternative to necessary services, services that were usually very expensive (O'Neil, 2016) (Dörr, 2016). Under the premise that humans cost substantially more, they offered Artificial Intelligent agents that could do the same job -if not better- more efficiently and more importantly, cheaper. At the same time, AI agents were much faster, did not take breaks or parental leaves, did not fall ill, or have any right that a human worker would do.

The problem is that the way Artificial Intelligent agents operate at the moment, they are not applicable for quality usage; rather, they are useful for mass-produced services only (O'Neil, 2016; Dörr, 2016). A robot is more efficient than a human in a production line factory where the job is easy to become automated. On the contrary, a judge, a school evaluator, and a banker who determines who gets a mortgage and who does not are occupations much harder to replace with algorithms. They are much more complex, rely on experience as much as critical thinking, and ethical principles too.

This is apparent in high-brow firms, like high quality law firms. In these companies, even tasks that are very easily automated, like the sorting out of CV's are being handled by human recruiters and not algorithms (O'Neil, 2016). This argument proves that Artificial Intelligent agents at the moment are only used by companies on tight budgets. Bad as it may be, the case is that these algorithms are not yet ready to be put into commercial use. The fact that this is a low budget solution proves the inefficiency of said algorithms. Had they been delivering exactly what their developers promise, successful companies would be the first to use them for their benefit. Instead, humans are trusted where agents are not, evidencing that algorithms are not yet to be trusted with anything but automated work.

5. Biased AI

In biology, scientists agree that the offspring of a certain individual inherit some of its characteristics. In computer science the same trend can be observed. Developers, even if unwillingly, put into their code norms and trends that are implicitly within their thinking. In other words, developers try to program unbiased algorithms, only to find out that it is highly
unlikely because these biases exist deep within their minds, where the actual code comes from. This notion has been termed *Engineered Inequality* by scholars, and has been a central theme of social scientists’ studies within Artificial Intelligence production (Benjamin, 2019; Noble, 2018).

It is important to note here that there might be various reasons for *Engineered Inequality* to occur. The most obvious, of course is that coders carry with them biases from their background, biases formed throughout their lives and are difficult to remove because they are not easy to spot. But there are other reasons. The fact that a programming team consists of a wide male majority might play a role in gender discrimination in algorithms. Again, developers do not willingly put certain genders under scrutiny, but tiny bits of their code might favor one gender over another; this problem might occur just because not a lot of effort has been put to distinguish it.

Other problems might be the ethnicities of programmers, as most of them come from countries with conservativism as a general background, or even the scope of the algorithm. In certain cases, algorithms have been made assumptions about a person’s economic status depending on their zip code. The argument here is that since some neighborhoods are considered poor, if one inhabits such a zone must be of a lower economic status than an inhabitant of the most expensive part of the city. While logical, this example sheds light to the biases programmers put in their code. Even if the above premise is accepted, identifying one as of lower economic status should not diminish her score for a mortgage. It might just be the case that she lives with her parents or rents an affordable house at the moment in order to save up enough money to pay off her mortgage.

6. Surveillance

Several reports have made apparent the fact that Artificial Intelligence agents have been used in surveillance cameras for facial recognition (BBC, 2020; Gayle, 2020; Valentino-DeVries, 2020; Ziady, 2020). Some cities, like London and Hong Kong had massive marches against this phenomenon (Potts, 2015; Mozur, 2019). While surveillance is an issue that should not be accepted light-heartedly, there is an even deeper notion within it that is even more problematic; inequality in the form of surveillance, a term named *Coded Exposure* (Benjamin, 2019). As described above, Artificial Intelligence agents need massive amounts of datasets to train, in order to function properly. In another argument above, I analyzed how certain communities of people are targeted extensively by Big Tech companies (usually the wealthier) while others remain marginalized. This leads to uneven mined datasets, as most of the information will come from people in the first category. These datasets are then used to train AI agents, creating more biases. *Coded Exposure* refers to the fact that even in gray-areas, such as surveillance, algorithmic biases are apparent (Benjamin, 2019).

This might seem trivial on the surface, but it is rather problematic. Facial recognition algorithms are used more and more as technology evolves and our societies incorporate it in
their routines. By the lack of data from the marginalized groups, said groups become even more marginalized as newer technologies do not incorporate them in their agendas. Steve Lohr wrote in the New York Times that “facial recognition is accurate if you are a white guy” (Lohr, 2018). More and more services shift from passwords to biometrics for security. Creating an account using facial recognition or even unlocking their phone might prove a difficult task for non-whites, as the algorithms are not “smart” enough to detect them. This is but one example of a larger problem; even in facial recognition, which has been framed as invasive and undesirable, algorithmic discrimination exists.

There is, however, another ethical issue arising which is rather deontological. Surveillance, in a sense, breaks moral laws on multiple grounds. In a more Kantian way, the universal norm of individual freedom is being violated by constant surveillance. People have the right to withhold some information solely for themselves; secrets are secrets for a reason. By using intrusive technology, thus, not only it is unethical from a consequentialist perspective, as analyzed above, but it goes against deontology as well, rendering surveillance unethical on multiple levels.

7. Corporatization of public Knowledge

While this argument has been briefly touched upon above, some more depth is due; that is because the future of searching does not look bright (Noble, 2018). As previously argued, some content will end up on the top side of the search results while other will be buried under a mountain of other content, making it almost invisible for the average user. The reverse, however, is also true. Noble argues that the content on the upper side of the results are going to be clicked more, thereby increasing the difference between them and the lower-result content. What this means is that as the searches scale up, instead of the algorithms mixing the results and trying to end up with the perfect outcome, they encourage this fight for the top spots even further (Noble, 2018). If one’s content does not end up at the top of the page from the very beginning, the situation is almost impossible to alter.

This, of course, is not only an issue with Google’s search algorithm but with every commercial Artificial Intelligence agent. However, for Noble, corporatization of public knowledge is problematic for two additional reasons. As Marcus and Davis informed us previously, echo chambers are very easy to be created within Artificial Intelligence agents. Combined with filter bubbles -a term mostly used for social media, where one tends to interact mostly with people sharing her worldview, therefore living in a bubble because she fails to see the complexity of the world- Artificial Intelligence agents are not only allowing, but instead promoting hate speech and hate crimes (Noble, 2018). Some social media preyed upon the “marginalization” of hate speech from Big Tech platforms and created their own communities of followers and believers in racist, elitist, and other hate speech, enhanced by the formulation of big, thick filter bubbles.
The second reason stems from the concept of “never forgetting”. This is a term that was devised by Noble to claim that once data are generated and stored in a dataset belonging to a Big Tech company, the user whose data have been collected has no legal or actual right over them (Noble, 2018). After a big backlash and massive campaigns where people demanded the right to be forgotten if they so desired, Big Tech companies agreed to unlink the profiling of users with physical entities. In other words, the profile that the company created would still exist, but it would not be linked to its user. Rather, it would still be prone to exploitation or an advertising target, but not as a certain person but as a certain profile with X and Y characteristics that have been collected and attached to this profile (Habib, et al., 2019). Noble thinks that, if so users desired, they should delete all of their data immediately, because it is rather stressful to live in a society where nothing is ever forgotten; one needs only to search the archives and find data that its owner wishes to be forgotten (Noble, 2018).

The next argument, in relation to the issues of corporatization of public knowledge, is a rather economic one. Under capitalism, free market thrives. As competition over a certain field increases, consumers are benefitted as different vendors will try to make their product or service more appealing for them. The reverse is also true. As competition is scarce, consumers are rather exploited, especially if the product at hand is necessary. When it comes to Artificial Intelligence agents, the market is surprisingly narrow. While many companies claim that they use Artificial Intelligence in many sectors, James Vincent informs us that “Forty percent of AI startups in Europe don’t actually use AI” (Vincent, 2019). The ones that are and they show potential are being bought by Big Tech companies, as it happened with DeepMind and its purchase by Google (Gibbs, 2014). This creates a very narrow oligopoly, instead of free market competition idealized by neoliberals, which creates further issues with choices and monopolistic practices. In other words, taking the example of search engines, by relying on Google and Microsoft as the only two viable players in the market, consumers not only are not benefited, but potentially highly exploited by various monopolistic practices performed by the vendors (Noble, 2018).

Lastly, it is important to note here that the problem of corporatization of information existed in the past; from librarians cataloguing content to the assimilation of knowledge itself (Noble, 2018). It would be a stretch to argue that this is a novel issue. However, within the Digital Era, this problem happens on an unprecedented scale. The mundane task of gathering and distributing information has been assigned to Artificial Intelligence agents which are way more efficient than medieval monks or librarians in Alexandria. The problem is that societies have to sacrifice something in order to gain efficiency; and most of the times, what is being sacrificed is the right to withdraw equally any kind of information a user is seeking. The fact that the ads on my Facebook newsfeed are advocating Artificial Intelligence is understandable, since this subject is of high interest for me, but it is inexcusable that on a similar newsfeed my friend gets ads arguing that the earth is flat. I have no access to the latter ads, as my friend has no access to the first ones, something which as I have shown, is problematic on various levels.
8. Universal Linguistic Need

Kerns and Roth argued that a convincing definition of abstract and loaded terms is yet to be holistically approved by social scientists around the world (Kerns & Roth, 2019). Indeed, labels such as democracy, fairness, ethics, and morality are just some of those terms in need of concrete definitions. It is impossible to move forward without clearly defining those -and other- terms, as developers need to have a clear understanding of what they should strive for (Kerns & Roth, 2019). To be fair to them, creating Artificial Intelligence agents are notoriously difficult to accomplish; by complicating matters even further, not only are social scientists not helping, but risk of being isolated by computer scientists as needlessly alarmists, or arguing without having a solution.

Ethics play a big role on trust, as it is highly unlikely for consumers and citizens to trust algorithms that either have no ethics or a distorted version of ethics in their core (Marcus & Davis, 2019). While striving for an objective definition of ethics might be just a pipedream, it should not be left unresearched. It is rather essential for the advancement of technology. Developers have evolved to the stage where they have to incorporate actual thinking inside their algorithms, and not simply mathematics. But to do so, social scientists should live up to the challenge and come up with a definition that will satisfy most -if not all- developers and consumers.

While O'Neil believes that fairness is not something that we can algorithmically achieve, I have to disagree. Instead of using actual fair thinking, Artificial Intelligence agents use proxies that unfortunately only correlate specific data to actions and vice versa (O'Neil, 2016). While this is true, the reason for that is that there is no scientific consensus as to what fairness truly means. If this milestone is achieved, the way that this will be programmed into algorithms is something that needs further thinking but in the future.

O'Neil also claims that by using Artificial Intelligence we justify through “technological and scientific methods” biases and flawed models that prey on the poor. In other words, this exploitation, not only is it not disagreeable with Big Tech, but it is rather a sophisticated way of marginalizing and discriminating without having repercussions. Again, Big Tech relies on the ambiguity of the terms fairness, ethics, morality and the rest in order to -maybe- promote their secret agendas. By addressing this philosophical issue now, not only can developers finally move forward to the next level of Artificial Intelligence, Big Tech will lose its biggest steppingstone yet; judicial indecision because of lack of clarity on the terms under scrutiny.

9. Artificial Intelligence as a tool

I would like to move the discussion now from what Artificial Intelligence is not good at and should not be used for to how it has been claimed to actually help society. Indeed, scholars agree that, if AI agents are not overestimated and be put into the right spots, they can work miracles (Diakopoulos, 2013; Diakopoulos, Paving the Human- Centered Future of
Artificial Intelligence + Journalism, 2019; Diakopoulos, 2020). It is, though, trickier than this simplistic notion. For most researchers, while Artificial Intelligence is not competent or robust enough to take over human workers, it is one of the greatest tools they have yet; a tool that in combination with human actors can enhance humanity’s ability to perform various tasks. But this is all we can expect from those agents, at least in their contemporary form.

Much like a hammer where one can use it to build a house or commit a murder, one can use Artificial Intelligence for both good and bad; algorithms are not ethical, nor should they be (Kerns & Roth, 2019). As long as people are the ones behind the wheel all the time, the situation is much harder to get out of hand. It is important, of course, to keep in mind that as the abstraction level of each tool is elevated, so should be our caution. Using a much more complex tool might prove helpful if used correctly but disastrous otherwise. Human actors should be able to understand their tools completely, lest they risk endangering their work and the other people involved.

This idea of AI helping humanity if viewed as simply a tool is a central notion for Broussard as well. For her, the third wave of Artificial Intelligence - where machines work together with humans - should be humanity’s next goal and top priority (Broussard, 2018). An example she gives is a combination of an AI agent performing calculations of moves in chess, therefore allowing more time for their human operator to devise strategies. As stated in the Introduction, Deep Blue managed to defeat Gary Kasparov in their 1 vs 1 match. However, when Kasparov was teamed up with an AI agent to assist him in his matchup against his mechanical opponent, team Kasparov proved to be much better. This combination of human-AI is called Centaur Chess and is now a standard form of chess playing.

This example proves Broussard’s argument that combining humans with Artificial Intelligence is indeed something desirable, something that can relieve society of certain repercussions that would be plausible, had AI agents been acting alone and unsupervised. Marcus and Davis state that Artificial Intelligence is not magic; it is just algorithms that are useful in some situations, useless in others (Marcus & Davis, 2019). It is common sense to use them where the environment renders them effective and when they are not, humans should take over. O’ Neil and Fry both claimed that while scholars focus on the worst cases of AI, it is a very helpful tool that can upgrade contemporary societies and make the quest for seeking knowledge easier and more effective. It is important, though, to know their limitations (O’Neil, 2016; Fry, 2018).

Finally, Marcus and Davis claim that AI agents should not try to imitate the human brain; this is not only unattainable but also undesirable (Marcus & Davis, 2019). This argument has core similarities with the early human attempts at flight. Those first thoughts were developed after observing how birds and bugs fly, and so humans decided to replicate them, with no success (Kline, 2000). It would take years and a leap of imagination to stop trying to mimic nature and abide by the mathematical laws of physics in order to create the first successful airplane (Smithsonian Institution, 2015). Similarly, researchers should try to incorporate the
Critique of Consequentialist ideas on AI

While an interesting idea, it is still constrained under the main issues concerning consequentialism; forecasting the outcome of a certain action is quite problematic, even with the help of immense computational power that AI agents provide. It is impossible to imagine all different scenarios that can unravel in real life when a certain decision is taken. But even if a moral agent could hypothetically forecast the totality of the outcomes, there is a second problem; proximity. Should the agent look only for temporal outcomes or include ones that will potentially arise in the future? The nuclear bombing in Hiroshima and Nagashaki for example, could be ethically justified by consequentialism, since American lives were spared, and the total human cost after the bombing was less than it would have been, had the war continued. What is not taken into account, though, is the fact that radiation and radioactive waste were quite an issue for a big number of people for many years after the actual detonation. In other words, predicting the outcome, even if an agent knows all the parameters, is a very risky notion.

More scholars have been arguing about similar models in their works, but Etzioni’s and Etzioni’s work was a combination of most of them. I will not argue that I have covered all of different arguments made about consequentialism, but I hope that the main ones are now well established. There has been some minor critique to those models, but consequentialist theory is hard to argue against, because it takes many different forms, and is quite abstract in its essence; what maximum happiness is for one party might be a total disaster for another. The biggest problem of utilitarianism, though, is the fact that is so elusive that it becomes dangerous to incorporate in official doctrines. Aristotle, criticizing Epicurus when he followed his own life of maximum pleasure, wrote that “this is the life of a beast” (Scarre, 1996).

Aristotelean Virtue Ethics

Although by far less popular than the previous two schools of thought, virtue ethics have started to gain some ground among social scientists. As discussed in Bowin’s “Aristotle Virtue Ethics”, this school of thought, instead of focusing on a set of moral rules (deontology) or the consequences of an action (consequentialism), it focuses on the action itself; in other words, if an action satisfies certain conditions, it is considered ethical (Bowin, 2020). This school of thought combines the two former ones into a more complex solution, requiring abstract thinking of multiple levels for its materiality. For Aristotle, the definition of virtue is any principle that derives from previous norms but also is beneficial for the actor (Bowin, 2020).
Bravery, for example, is considered normatively a virtue. While bravery is good, too much bravery is considered recklessness, therefore should be avoided. On the other hand, too little bravery is considered cowardness, again something that should be avoided. The reader should keep in mind that Aristotelian virtue ethics judge morality based on pre-existing values and the benefit or maleficence of an action for the actor or his/her community.

The connection between virtue ethics and the other two main ethical schools of thought did not go unnoticed from MacDonald and Beck-Dudley. They claimed that virtue ethics are deontological in their main sense, as they incorporate a normative approach on framing a term a virtue (MacDonald & Beck-Dudley, 1994). Without pre-existing values, the argument goes, it is hard to imagine how virtues could be established. There is not a clear approach as to how virtues are generated; rather, they rely on categorization available from a different perspective. Virtue ethics, though, do incorporate consequentialist ideas, because the final decision of a value turning into a virtue focuses on the consequence of that action (MacDonald & Beck-Dudley, 1994). Even though the action itself is under scrutiny in virtue ethics, the prism through which they are examined is their consequences. In the aforementioned example, recklessness is considered not a virtue because it will most likely get the actor injured or killed. So, even if the focus is on a different level on virtue ethics, both deontology and consequentialism are being present in their discussion.

Bench-Capon, in his article “Ethical Approaches and Autonomous Systems” uses the myth of “the ant and the grasshopper” and the biblical parable of the prodigal son as scenarios to test both major schools of thought, as well as virtue ethics in their applicability in Artificial Intelligence ethics (Bench-Capon, 2020). The questions were mainly two, focusing on the first myth, placing the AI agent in the grasshopper’s shoes in the first and the ant in the second: a) since it is summer, should the agent work or play? b) should the agent give some of the surplus to another agent that did not work?

In the consequentialist doctrine, where most of utilitarian ethics are incorporated, the agent should prioritize work over play in the first question, because this leads to an outcome of dependence to another agent (if play) or independence (if work) (Bench-Capon, 2020). In the second situation the agent should punish the dependent agent (the one who played) because this will educate the latter to not repeat the same mistakes. However, this outcome might prove detrimental for society in the long term as many agents will not survive. Rather, as the prodigal son parable teaches, forgiving once is not only ethical but desirable (Bench-Capon, 2020). Another important argument for Bench-Capon is that developers of AI agents should be held responsible, should an agent act against the law or immorally.

In the deontological doctrine, where absolutism is incorporated, the agent should prioritize work over play because this is a normative rule that should not be broken under any circumstances. In the second situation, the ethical agent should always punish the other one because the latter prioritized play over work, rendering the agent immoral. The same argument about the prodigal son applies in deontology as well. Developers should be held
completely accountable for the effectiveness of their models, because they are the ones imposing the normative values into the algorithms. Even though the reason behind every action is different, this example’s brilliance is that it shows that the outcome is the same. Irrespective of the model social scientists have chosen to analyze, research was stagnating because these schools of thought do not differ in this context.

Bench-Capon though wanted to show this stalemate in order to initiate virtue ethics into the discussion. In the first situation the working agent is considered ethical while the playing agent unethical, because working and gathering for a time of need is considered virtuous, while seeking pleasure without security is considered a vice (Bench-Capon, 2020). This is in direct lining with the previous two schools of thought, as well as with -arguably- most people’s perception of what is ethical in this situation. In the second situation, however, the working agent has different choices: it could give food to the playing agent therefore being altruistic (virtue), it could not give food to the playing agent therefore being just (virtue), or it could try to reason with the playing agent to take turns working and playing therefore being fair (virtue). In other words, instead of a certain path of instructions irrespectively of the scenario, virtue ethics allow for maneuverability, making it much more useful in real life scenarios. Unsurprisingly, the developers are, for Bench-Capon, again responsible for choosing a system under which these virtues and vices will result in the optimal solution (Bench-Capon, 2020).

Another scholar that examined virtue ethics in his AI ethical model is Mike Ananny. In his article “Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness” he stated that, much like this essay addresses, there are three major school of thought for AI ethics; Deontology, Consequentialism, and Virtue Ethics (Ananny, 2015). His model though, criticizes all of them and uses them at the same time to build up a model where a combination of all those works best. While not an advocate of Virtue Ethics per se, Ananny recognizes the importance of this school of thought as both influential and a steppingstone. In other words, by using virtue ethics explicitly as part of his proposed model he normatizes the ethical theory as not only acceptable but beneficial for an ethical discussion about AI.

Frank Pasquale in his article “Data-informed duties in development” advocates that regulations should not only restrict but guide ethical decision making in AI creation (2019). Much in accordance with the consequentialist approach, he proposes that AI developers should keep the endgame always in mind through his “standards of care”. These standards are much related to Google’s motto “don’t be evil” and they basically strive for win-win situations between developers and users. Pasquale’s belief is that the usage of flawed data can be harmful to not only users but to the agents as well, as they will underdeliver (Pasquale, 2019). Additionally, compensation might be due from the companies relying on flawed data for breach of trust. In the end, this consequentialist thinking will help the judiciary system to have a unified approach.
Critique of Aristotelian Virtue Ethics

However refreshing this idea might be to a field where similar patterns are continuously recycled, major limitations exist for Bench-Capon’s model. It would be unfair and unrealistic to hold developers responsible for every possible scenario that might occur in real life. It is highly unlikely that every possible situation could be programmed inside the algorithms. Even if the notion that more abstract coding might be available is accepted, and therefore algorithms will have a general instruction rulebook to operate from, this circumvents the usefulness of virtue ethics - the analytical thinking of every different situation - because the agents will follow a certain rulebook. In other words, virtue ethics will not be much better than deontology and consequentialism if they are not autonomous enough to judge each situation separately; something which is rather unlikely with the contemporary Artificial Intelligence technology. It is true that AI gets more sophisticated. It is also true that with more complex equations, harder problems have been solved, while others are being solved at the moment. This, however, is a problem of simplification, not abstraction. In order for Virtue ethics to work, each action should be judged separately. AI helps with combining more and more variables, arriving at complex conclusions. Virtue ethics are more deconstructivist in this sense; the more isolated an action is in order to be examined, the better. I don’t see how compiling data over data, as AI is doing at the moment, can help tackling this issue.

Such limitations do not turn Bench-Capon’s model useless but rather unsuccessful attempts to conceptualize an interesting idea. In the Discussion section I will propose a similar model, taking into account current limitations and proposing a novel solution based on Bench-Capon’s idea. It should be noted that I will depart from Aristotelian virtue ethics, as to my mind, they are very similar to both deontology and consequentialism. There are direct relations to all three major ethical school of thoughts, as their founders or major influencers have been ancient Greek philosophers. The next section is dedicated to these interrelations in order to make clear that a complete shift in how social scientists view ethics is needed if they want to get away from contemporary problems concerning AI ethics. Like a maze, there are many different paths, but they all lead to the same endpoint. Like a maze, one can lost herself completely in an infinitum of dead ends, only to realize that where she ended up is where multiple scholars arrived well in advance. Only by leaving said maze behind can she explore new trails.

Solutions

Lastly, after an extensive analysis of many arguments about issues relating ethics in Artificial Intelligence, it is due time to unfold some solutions that have been proposed by scholars. These solutions can be condensed into two general arguments, something which I think is helpful for this project because of space limitations. The two arguments concern regulations around Artificial Intelligence development. Most of the scholars agree, even though there are some who believe that regulation is not the optimal solution.
The problems associated with Artificial Intelligence are not computer problems in their essence but rather human. For this reason, human regulation and pre-thinking policies should exist (Broussard, 2018). It is quite apparent that Broussard trusts regulators and policymakers to be more concrete and specific than social scientists and philosophers when it comes to policy making. Otherwise, it would just be just another failing attempt at grasping something elusive; trying to imprison an idea is optimistic, to say the least.

Noble agrees with Broussard that further public policies are needed, as well as further regulation in order to protect democratic rights from neoliberal corporatization (Noble, 2018). In the end, basic economics preach that if the free market is not able to find equilibrium, the state should intervene in order to artificially create equality among the parties, lest one would be constantly exploited. It is hard to disagree with the second premise, as she has convincingly argued about the arising issues of corporatization of knowledge (see the subsection dedicated to this issue). The first premise, however, the fact that governments and regulators will be able to solve the problem is a bit ambiguous; it doesn’t really provide a solution, but rather throws the ball to another field, hoping that they might be able to help.

Kerns and Roth disagree with both Broussard and Noble. They think that external regulation, with the example of GDPR does not work as ideally as it should for two main reasons. The first one is that it furthers complicate problems (Kerns & Roth, 2019). By paying fines, Big Tech companies have a license to continue using their discriminatory practices. If the price that they have to pay is monetary and not enough to shake them, they are going to continue their problematic ways and pay fines to equal out the damage that they have done; something ludicrous in its concept. The second problem is that it provides a steppingstone for companies to find holes in the system. A very recent example is GDPR’s revision of article 17 (formerly article 13) and the backlash it initiated. It gave room to YouTube to publicly announce that “the final version of the EU Copyright Directive is an improvement, but we remain concerned. Article 17 (formerly Article 13) could still have unintended consequences that may harm Europe’s creative and digital economy” (YouTube, 2019). In other words, YouTube argued that regulation might be harmful for creative and digital economy, a notion which Google (the company behind YouTube) has been proud to disrupt!

Gry Hasselbach wrote in her article “Making sense of Data ethics: The powers behind the data ethics debate in European policymaking” that GDPR is currently in deep waters, as the EU did not manage to make clear its intentions from the beginning (Hasselbalch, 2019). There is an uncertainty that has not yet been addressed but Hasselbach is optimistic that it will happen soon. Additionally, she tackled the thorny issue about GDPR’s claim “that we [GDPR] recognise data ethics policy initiatives as open-ended spaces of negotiation among different interest groups that seek to guide the cultural definition of ‘data ethics’, with complex power relations exercised via cultural positioning” (Hasselbalch, 2019). The view, thus, that AI is problematic and needs fixing is apparent for the European Union, even though a clear definition of what these data ethics are has not yet been established.
Creating space for companies to obscure reality and strengthen their market position is the exact opposite than what Broussard and Noble wish to accomplish with their ideas. Unfortunately, the most realistic scenario is that as regulations pile up, the idea of digital liberators will be further instilled to Big Tech companies. Even though this has been proven very far from the truth based on my previous argumentation, turning huge corporations into people’s saviours is something that should be highly avoided. It seems, thus, that more regulation is not the best way forward, at least not in the current state of events.

Another solution, which has been proposed by Kerns and Roth is quite bizarre; more algorithms. Algorithms that will take into account not only the optimal solution to a problem, but the optimal solution AND fairness, ethics, morality, etc. (Kerns & Roth, 2019). In other words, what they propose seems like fighting fire with fire. At the state Artificial Intelligence is at the moment, had an algorithm needed to take into account multiple variables at the same time would render it sub-optimal. As a consequence, developers will not agree with downgrading their own agents because they will lose their competitive advantage. At the same time, consumers might not agree using sub-optimal Artificial Intelligence, since a current level has already been established. It is very hard to go back in technological advancements. Imagine if a regulation came out that forbade the use of smartphones or making real the breaking up of the internet. People would strongly react because a certain level of dependency has already been achieved.

Kerns and Roth propose to test them in controlled environments, train them adequately, and through natural selection find out which work better and why. Artificial Intelligence usage is not one-size-fits-all (Kerns & Roth, 2019). Consumers and developers might decide that the level of “fairness” when it comes to movie recommendation algorithms should be much more loose than that of agents involved with justice delivery (O’Neil, 2016). This is an interesting approach. It hasn’t come up with many researchers even though it is an argument worth mentioning. Social scientists tend to see Artificial Intelligence as one entity while it is not; rather, it is many segmented pieces falling under the same umbrella term, following different paths and resulting in different outcomes. This solution claim Kerns and Roth is a mix of macroeconomics and game theory, as it interacts with the notion of win-win situation in a truly free market.

The only problem that has not been addressed by either Kerns and Roth or O’Neil is the fact that they take the loaded terms for granted. While it is very convenient to think of developers’ and consumers’ repercussions as to how they will react to this solution, the problem at hand exist in the core idea itself. Namely, in order for an algorithm to be optimal and fair at the same time, fairness has to be defined first. As I analysed in the subsection universal linguistic need, what these scholars address is the second stage, that of how it can be incorporated into the algorithms themselves. The first and immediate issue is that of a clear definition, something which scientific libraries lack at the moment, something which has
to be formally addressed soon, and something which is going to take up the rest of this project, beginning with the next section of the literature review, Ethics.

**Common Roots in Ethical Frameworks**

What I intend to do in this section is to link the three classical philosophers (Socrates, Plato, Aristotle) to deontology and consequentialism, the two major schools of thought that have been presented in the previous sections as the main influencers of Artificial Intelligence Ethics. A relation with virtue ethics - the third school I analyzed above - is not needed as its founder was Aristotle himself. That is important for two reasons. Firstly, it shows that novel arguments about modern problems are not as novel as they seem, since they are based on ancient philosophical ideas. Secondly, if some ancient Greek philosophy is used, I don’t see why it should be limited to just a small portion and not the totality of it. After that, I will have some further comments about the use of Logic, and how it has differed from Ancient Greece to the 21st century. Finally, I will briefly introduce a new voice to the discussion, Protagoras, which will be my major influence for the following sections.

Contemporary ethical theories, as one moves further and further back sees, have common roots in the classical and world famous ancient Greek philosophers Plato, Aristotle, and their teacher Socrates. The latter did not leave any written texts behind, but readers can get a sense of Socrates’ influence through the work of Plato and Aristotle. Dreyfus and Dreyfus went so far to argue that even the term Artificial Intelligence began with Socrates and (after) with Plato, when they argued that every human notion can be put down to sentence and logical statements (Dreyfus & Dreyfus, Mind Over Machine, 1986). For them, this has been the triggering thought for all of contemporary research and development on the AI field, because it took form into classical literature and dominated our culture and doctrines. In other words, through their authenticity, the notion that every human characteristic can be explained through logic was embedded into contemporary thinking. Being unable to root this statement all the way back to its origin, scientists took it for granted and began developing Artificial Intelligence agents based on a seemingly logical procedure.

Dreyfus, digging on the previous notion, argued also that Plato went even further to argue that all morality should be epistemological; it should be explained and put into definite, normative terms (Dreyfus, What Computers can’t do: a Critique of Artificial Reason, 1972). To put it in different terms, if something cannot be explained logically and mathematically it is rather intuition than science. While this might seem like an interesting idea for a 5th C BCE philosopher, the same cannot be applied for contemporary scholars. Trying to see everything through a prism of Logic, one that will be able to explain everything with mathematics and objective definitions is dangerously simplistic. Real life situations are endless and quite diverse; even if philosophers and scientists manage to put them all into different categories, chances are that said categories will be very abstract and therefore useless. Objectively
defining the Earth as an almost spherical object, a bit compressed in its poles is something possible. Using the same method to define ethics or morality is mind boggling.

Plato’s and Aristotle’s theorems of using Logic to define humanity, though, have been extremely popular ideas, not only because they were part of societal narrative. Logic is an extremely efficient tool that has been used for several projects so far, some of them having a huge impact in humanity’s future. The rise of the personal computer, the internet, smartphones, or generally any device with a microchip performing calculations owes its existence to Alan Turing, Bertrand Russel, and as a consequence, Plato and Aristotle (Huth & Ryan, 2004). Artificial Intelligence development could, obviously, not be left out of this discussion. Aristotle’s logic is excellent for producing AI according to Reva (2018) because this way everything can be turned into binary code and programmed into algorithms. These algorithms, then, can be loaded into mechanical parts, and the rise of robots is unavoidable.

This procedure though has severe limitations. The first is that, as Reva argued, Aristotelian Logic is not useful at explaining human understanding (Reva, 2018). While using Logic as a tool to program algorithms might seem an excellent fit, there is no clear connection between explaining human ethics and translating them into numbers. In other words, Aristotelean Logic fails the test that Plato proposed and presented before; if something cannot be explained logically and mathematically it is rather intuition than science. A case might be here that since ethics cannot be explained through Logic, ethics are mere intuition based on the definition above. While this might be true, ethics are accepted as a much-needed branch of philosophy in contemporary society (Toulmin, 1960), so a slight tweaking of the definition is needed: if Logic cannot explain that which exists, there is a flaw in said Logic.

A second important limitation, in close proximity with the previous one, is the fact that Artificial Intelligence agents have no morality, because they can have no morality; logic cannot decipher emotions, therefore they cannot be programmed (Reva, 2018). As a result, they allow for no plasticity, because one expects that, unlike humans, given the same scenario multiple times without feedback, the machine will always perform the same action. Since there are no ethics to be programmed into these algorithms, the outcome should always be the same independent of the ethical framework and the minor differences in the background. The founders of AI thought that Logic would be enough (Perlovsky, 2007). However, as proven mathematically, there is an imperfection of the human mind to understand Logic (Perlovsky, 2007). This happens because, for Lucas-Pernose, the human mind is not a formal Logic system (Perlovsky, 2013). Put differently, what Perlovsky claimed is that Logic is not enough to understand the human mind, because the human mind is not primarily logical in the first place. Overran by emotions, norms, prejudices, and experiences, human brains are much more complex than organic calculators. Dreyfus argued that intelligence is not defined as processing speed in general application in Plato’s term, but rather efficient in a single task (Dreyfus, 1972). This notion is apparent with AI agents, as they are raised to outperform humans in certain tasks but remain oblivious of others.
Deontology

Plato’s and Aristotle’s ideas though had not only influenced contemporary ethical debates. Deontology as a discipline, even though its main contributor has been Immanuel Kant, he has stood in the classical ancient Greek philosophers’ shoulders in order to shape his arguments. Korsgaard wrote that “Platonic justice and Kant’s categorical imperatives are shown to be normative standards for action because they are principles of self-constitution” (Korsgaard, 1999). Seidel claimed that there is, for Plato, an objective reality (Seidel, 1991). This is highly correlated with the deontological discourse, because in the center of deontology’s normative beliefs there is objective reality; absolute truth, that has to be followed to the letter, lest one is immoral.

Central to this objective reality of Plato, Aristotle, and Kant, is the core of epistemological knowledge as scientists currently define it (Seidel, 1991). Observation, hypothesis, testing, and conclusion are the standard procedure that every scientist accepts at the time of this essay’s writing. Interestingly enough, this model implicitly accepts that objective truth exists; a truth that is well hidden. An objective truth which scientists are called to uncover. It is no wonder that theories of objectivity have conquered the scientific communities; they are the most convenient to promote the scientific ideals. In the end, Plato’s notions are programmable, since they are based on Logic and contemporary consensus on the scientific principles (Seidel, 1991).

Tasić tried to get away from this notion, arguing that for hundreds of years, objectivity was ruling the world; from Socrates, to Einstein, to Hawking (Tasić, 2015). Scientific practices had little in common throughout the ages, but a strong belief in objective truth was one exception; an exception that is now under heavy fire with the emergence of quantum physics. Thus, humanity enters a new era, that of relativity and subjectivism (Tasić, 2015). Tasić raises an important point here, as quantum physics are a complete breakthrough from the classical Newtonian and General relativity theories. Even though the last two have been very helpful in order to understand not only our planet but parts of the universe as well, by entering this new era of quantum physics, this notion that the truth is hidden under layers of protection, waiting for scientists to uncover it, is considered obsolete. As in quantum physics a particle can occupy multiple positions at the same time, so does ethics, morality, and truth; they turn out to be relative, subjective, and different (dependent on one’s perspective).

Consequentialism

Similarly, Plato and Aristotle shaped consequentialist theories as well. Scarre explicitly stated that utilitarianism has been greatly shaped by Aristotle and Epicurus (Scarre, 1996). While they were not true utilitarians since the term did not exist during their time, as well as their philosophies differed on multiple levels, their teachings helped to shape the movement (Scarre, 1996). Aristotelian ethics, as I argued previously, do not focus on the consequences of an action but rather on the action itself, but this is not to say that Aristotle did not shape
these beliefs based on the outcome of an action. In other words, Bentham and Mills, inspired by Aristotle’s ideas, tried to polish virtue ethics and make it relevant to their societal needs. It is apparent, thus, that Plato and mainly Aristotle have been the pillars on whose height the whole utilitarian movement began (Scarre, 1996).

But there is a second connection between utilitarianism and the ancient Greek classical philosophers. Mills’ idea of happiness - the one that should always be considered to be at the maximum levels possible when taking an action if one wants to be ethical - are within very close proximity to Aristotle’s idea of a virtue (Scarre, 1996). It should be noted that Mills did not argue for pleasures of the flesh or hedonism as did Epicurus, because that would be “the life of a beast” in Plato’s words (Scarre, 1996). What, however, is promoted instead as the ultimate value is happiness as an ideal. A happiness that can take multiple forms depending on the occasion, a happiness that can vary between one deserting the battlefield to embrace one’s lover again, to being courageous in battle and not allowing one’s brothers in arms to get captured or killed. The link between Aristotle’s virtue ethics is as direct as it could be, with the exception that Aristotle was more deontological in his choosing of said virtues.

**Final thoughts**

Even the definition of knowledge itself that is currently in use derives from Plato and Aristotle and has three premises: a) a person believes something to be true, b) this belief is actually true, c) the person is justified for believing it (Groumpos, 2018). These premises, much to Aristotelian style, are very easy to be implemented into Artificial Intelligence agents. Surely enough, as algorithms are powered by Logic, they can work ideally with objective definitions of reality as I argued above. Without much external interference, by having an objective truth embedded in the algorithms, the AI agents can calculate the possible outcomes of an action and choose the one that best fits their norms. The problem is, again, that the evaluation of the second (b) premise is difficult. In order to claim that a certain belief claimed is true or false, an a priori consensus must be reached beforehand, lest the evaluation will be impossible. Even though subjectivity is not universally acclaimed triumphant over objectivity (as that would render subjectivity objective truth), in some situations objective truth is possible and impossible in others. Generally speaking, human interactions with the natural environment tend to be much more objectively measured rather than interactions between humans. It is, however, in the latter category that Artificial Intelligence Ethics falls under and trying to force objectivity where it does not belong offers no benefits whatsoever.

A last argument I would like to present comes from Perlovsky who reminded the scientific community that Logic, for Aristotle, was a tool for politics and not for science (Perlovsky, 2007). Interestingly enough, this argument has not been very popular among scholars, even though it seems very important to me. Since Aristotle “invented” Logic as a rhetorical instrument, one might say that my previous claim about human to human interactions are
not applicable to Logic is flawed. If Logic was used for the art of public speaking, the most basic of human interactions, not only should it be perfect for Artificial Intelligence ethics, but the conclusions should be obvious by now. Since this is not the case, allow me to offer an alternative explanation to this issue.

Logic, even though it was invented as a rhetorical instrument, it was used as a mean to argue about quantifiable data and scientific knowledge. The reader should keep in mind that rhetoric was used in the conventions, where strategical decisions were taken, and issues of state discussed. This is very fertile ground for Aristotelian Logic, because a) it uses quantitative data to arrive to conclusions, and b) uses a mathematical model to argue in natural language. Let me take the famous example of the two true premises lead to a true conclusion: premise 1: Socrates is mortal, premise 2: Socrates is a man, conclusion: therefore, all men are mortal. This is a perfect representation of Aristotelian Logic used for rhetoric, allowing for conclusive generalizations from smaller premises.

While it is very easy to think of this model as a tool to argue in a debate, the correlation in ethics is a bit less apparent, if at all there. That is because Logic is, again, standing in objective truths -in this case Socrates’ mortality and his human characteristics- while in ethics this luxury is rare. Most of the times, conclusions have to be abstractly made in order to have some foundations for the development of the theory. This is the case with virtue ethics as well, as Aristotle decided normatively some values that are likeable and some vices that should be avoided. Since Logic needs objective data in order to work -much like algorithms- its rhetorical use is unmistakably relevant, but the same cannot be argued about its applicability to Artificial Intelligence Ethics.

Having now sufficiently explained the influence of the ancient Greek classical philosophers in contemporary literature as a whole, I would like to introduce a much-marginalized voice from that era, that I think might hold the key to this discussion; Protagoras. Since contemporary scholars embed in their argumentation most of the 5th C philosophers of ancient Greece and the result is a grand stalemate, I thought to use Protagoras’ doctrines and teaching in order to bring fresh material to the discussion. It would be irrational, though, to believe that Protagoras’ arguments concern the Artificial Intelligence ethics discussion per se. However, much like other scholars used Plato and Aristotle -willingly or not- as a prism through which they explained their ideas, I intend to do the same with a sophist that has been mistakenly left out of the discussion. In the section that follows, I propose my solution to the problem through Protagoras’ ideas.

Discussion

In this section I am going to argue that relativism might be a fitting key to unlock Artificial Intelligence Ethics, even though so far it has never been considered as an option. Contrasting the objective models that dominate the contemporary discussion, I will propose
Protagorean subjectivism as a theoretical solution to this ethical stagnation. Then, based on this, I will propose an ethical model that relies on the personalization of AI agents. In a way, I propose to open the ethical shaping to the users instead of the developers. Lastly, I give rise to a famous Protagorean saying (two logos fragment) that further enriches the argumentation in favor of my model.

**Relativism over Objectivism**

Protagoras was fond of relativism, even though the terms “subjective” and “objective” did not exist -in their modern sense- in ancient Greece (Schiappa, 2003). Heavily disagreeing on this subject with Socrates and his followers, Plato described the debate between the two men in one of his dialogues named after the person under scrutiny, *Protagoras* (Plato, 1996). Based on the analysis in the preview sections it is quite apparent that the object of disagreement with Socrates was ethics as a subjective or an objective notion. While the latter tried to argue that objective truth is indeed embedded into the natural world, Protagoras claimed that humans are the measure of everything; if something cannot be put into human perspective, it does not exist7 (Schiappa, 2003). This fragment, called “the human fragment” for reference is viewed as an answer to Parmenides through the discussion with Socrates (Schiappa, 2003). Parmenides thought, much like Socrates and his followers, in binary terms; “A is B” or “A is not B”. Protagoras believed that “A is B for C” or “A is not B for C” (Schiappa, 2003).

This relativistic approach makes clear that C is the judge, the measure of A being or not being B. In other words, something can exist in multiple forms at the same time, as different people interact differently with their environment. For example, the same wind might be cold for one and warm for another; therefore, the wind is both warm and cold at the same time. This might seem like playing with paradoxes, but unlike Zeno -who was famous for finding mathematical and linguistic paradoxes- Protagoras is providing an example from the natural world. That is to say, the objective fact is that the wind blows, but if it is cold or warm wind cannot be objectively measured, lest an arbitrary norm is put, say, any temperature below X degrees is considered cold. But even if this last input is accepted, the fact that the objective definition is cold or warm does not alter the fact that different people might still be feeling it quite differently.

This argument is in complete alignment with Artificial Intelligence ethics. All in all, the current discussion as explained in the Literature Review is if an objective definition can be approached, and if it can, how can it embrace different views under the same concept. Taken Protagoras’ doctrine, trying to have objectivity in ethics is a fallacy. Even if one definition is

7 This is an unofficial translation by the author. The original piece is “πάντων χρημάτων μέτρον ἐστίν ἄνθρωπος, τῶν μὲν ὄντων ως ἔστιν, τῶν δὲ οὐκ οὖν ὄντων ως οὐκ ἔστιν“.
indeed accepted, and that is a very big condescension, it will only manage to answer some people's concerns. In many cases, these concerns might be in direct contradiction with others, so it will be a game of power as to who will abstractly, and most of all arbitrarily, legitimize her standpoint over others. If history can teach a lesson, scholars have proven quite explicitly that this is something which will only lead to more inequality. In other words, Artificial Intelligence ethics, the tool which is supposed to equalize the scale between the very powerful few, the many oppressed, and the excluded or marginalized few is going to be turned into a legal and moral license to further this injustice.

Let me get back now, to Protagoras’ example about the cold or warm wind. I accepted before the objective statement that the only certain premise in the example is the fact that the wind blows. Protagoras though did not think that even this is an objective fact. For him, something can both exist and not exist at the same time (Schiappa, 2003). In order to exist, the object has to be measured or experienced by humans. A second example he provided was the taste of a certain food. If it tastes sour for one and sweet for another, it is both sweet and sour. However, if the first person has not tasted it yet, then the sour effect of the food is not yet true, therefore making it solely sweet (Schiappa, 2003). To relate it to the previous example, if the wind blows but there is no one there to experience it, for Protagoras, the phenomenon does not exist.

While this concept seems quite difficult to grasp, as the wind blows irrespectively of one’s experience, it might get a bit easier to accept if thought through different terms. Since I stated that the wind blows, it is paradoxical to later say that it is not a fact that it did. In real life, though, this narrator does not exist. Schrödinger gave a much more realistic explanation with his famous “cat experiment” (Schrödinger, 1935). A cat in a closed box can be either alive or dead, but without peeking inside the box it is impossible to know which one is true; therefore, the cat is both alive and dead at the same time. If I extend this a bit further and do not give out the information that a cat is in the box, it would be impossible for the observer to know that a cat is there; therefore, even though an alive or dead cat exists in the box, for the observer it is none because she did not see it happen.

Even though this paradox was an explanation to quantum mechanics, this example fits beautifully into this discussion, as without knowing a priori if something happened, it is highly unlikely to believe it as true. Similarly, with Artificial Intelligence ethics, another problem occurs if an objective standpoint is selected. A potential committee responsible for choosing the definition might be oblivious to several marginalized issues and therefore not incorporating them in said definition, for lack of knowledge on their part will result in further marginalization. As Protagoras explained with the human fragment, people know only things that have been personally experienced or reported (therefore experienced from the reporter). It would be extremely difficult to argue for a total diverse committee so that every voice will have a chance of equally being heard. Relying on the data gathered so far is again
problematic because, as I argued in Literature Review, they are not collected on an equal basis.

**Subjectivism fits AI Ethics**

Going back to Protagoras’ relevance about subjectivism, I would like to argue in favor of relativism for Artificial Intelligence ethics. Irrespectively of the acceptance or not of Protagoras’ second remark, the fact remains that if something is not experienced, it is quite hard to be accepted, being objectively true or not. It would be quite difficult to promote objective values, because some people abide, trust, and believe in said values but others do not. I see no reason why an objective definition should be made in the first place. This seems like a bad idea on multiple levels. Firstly, it might be one more way for people in power to gain more power. Taking the inescapable fate of having to choose between one ethical theory over others, people in power might favor an ideological theory that best suits them. Taking an example by country, if most AI developers and stakeholders live in the global west, they might lobby their way to an ethical theory that will suit their ideas over those of an eastern country. Secondly, and very closely related to the previous point, it might be one more way for marginalized voices to be pushed even further aside. In a sense, not only will some people gain more power, some others might lose even the very loose grip they had previously. Thirdly, it might create unnecessary tensions to certain worldviews. Fourthly, since this is the beginning of a new era, this ethical framework will shape the way forward. On the other hand, by abiding to a subjective morality, not only are these issues not relevant, but certain positive outcomes emerge as well. A way forward and an alternative way of thinking about ethics are just two prominent examples, since the binary notion of “objectivity or nothing at all” is what governs contemporary western thinking.

Subjectivism, of course, cannot come without some basic reins. As Antonites claimed, relativism is acceptable, but it needs a base to be built upon (Antonites, 2005). He called for a “relative relativity”; Protagoras should be put on the one side, foundationalists on the other, and establish some middle ground (Antonites, 2005). This middle ground will accept relativism but will be built in some objective facts. Objectivist theories will be limited and as little as possible, but enough to balance subjectivism’s weak point. These foundationalist claims will concern universally accepted shared values, such as the value of human life. This seems like a good starting point because it would be irrational to delete research conducted for years by social scientists using objectivist theories. Rather, I am going to use just the best aspects of it and incorporate it into Protagoras’ ideas in order to argue for a different model of ethics. Richard Rorty proposed three different aspects of relativism. The first one is that every belief is as good as another. The second is that different justifications exist for different meanings, and the third that there are no clear definitions but only different descriptions based on one’s perception (Rorty, 1991). I remain skeptical about the first aspect as it makes it quite easy for total chaos to emerge. The latter two though, even though they have been discussed from Protagoras already, are concrete evidence that relativism, even though
marginalized, is not entirely misunderstood. While Rorty did not argue about Artificial Intelligence in his article, I can use his ideas and match them into the ethical discussion which this project concerns.

A final note before I proceed with my solution is important here. I am going to use Protagoras simply as a prism, through which I will propose my model. It cannot be denied that Protagoras has been wrong in some cases, even antithetical with his own teachings, as the norm in ancient Athens was to own several slaves and marginalize women and non-native populations. Similarly, some of his arguments on relativity might seem obsolete or too far-fetched compared with humanity’s progress two millennia after. However, given these flaws, Protagoras has been an inspiration for both the model below, as well as introducing the theory of relativism to me.

**Proposed Solution**

Based on Protagoras’ “human fragment” and its meaning behind it, it would be safe to argue for a subjective ethical framework that does not abide by normative rules, save very few as explained above through Antonites’ argument (Antonites, 2005). An ethical framework for Artificial Intelligence agents that would instead inherit its user’s morality and ethical perspective. In other words, a personalized agent that would not come with predetermined and predeveloped ethical standards but would encapsulate the worldview of its human agent. This discussion concerns of course the ANI (Artificial Narrow Intelligence) as with higher levels of abstraction and autonomy as is the case with AGI (Artificial General Intelligence) a different or a revised version of this proposal might be used.

Algorithms like these will be trained by each individual, much like kids are, in their younger years shaped from -mainly- their parents. Interestingly enough, since -in reality- a person’s ethics and morality are not the same in every application, the result might be that the same user, if using different algorithms for different aspects of her life, the AI agents might have different ethical frameworks depending on the feedback they receive from their human agent. In other words, diversity in AI might be achieved not by abstractly trying to think, forecast, and painstakingly program every agent by a small group of people but open it to the public and outsource ethical standards and unpredictable situations from a vast pool of users. This way, not only diversity is achieved much faster and more efficiently, but true diversity arises as well, since even marginalized voices can shape their AI agents however they want.

---

8 This way, there is indeed the risk of introducing the unwelcoming consequences of the internet culture, e.g. racist robots, etc. More on that will be explained later on, but the main premise is that no theory is waterproof. However, not every argument is of equal weight; it is my belief that the major issues should be addressed firstly.
Protagoras, through the human fragment, initiated a novel concept; getting away from Logic and binary thinking and trying to find alternative solutions. In the end, Logic can program computers. It is simplistic to think that much more complex organisms like the human brain, emotions, or ethical standpoints can be put into such normative terms. By personalizing each algorithm depending on its user, how morality can be turned to code is revolutionized. Not only is it simple enough for a machine to learn -provided constant feedback is given to it- but humans might feel responsible for their AI agents, since they will be part of them. In a way, it seems like parenting a kid that will make its agent’s life easier in its designated way. The responsibility for the agents’ decisions will fall, as is the case when parenting children, on the parents-users. This way, developers will be relieved of the pressure to create machines that can, potentially, create havoc and mass chaos, if extreme caution has not been taken. As I have argued previously, that pressure is unfair to them, since in order to have an objective ethical framework (if it were possible) embedded into AI agents, the algorithms are extremely complex. Minor mistakes are impossible to avoid even though the stakes are high. In this way, both complexity and unnecessary pressure are taken of them.

But there is one more thing that is taken off them; unprecedented power. By relying on an objective definition of ethical standards, developers could lobby their way into adding to the definition aspects for their own personal gain. Similarly, they could intentionally or not embed some minor adjustments to the programming of their AI agents in order to gain a leverage over competition. Or, they could simply game the whole process, as Ochigame has shown (Rodrigo Ochigame, 2019), and either delay the whole process indefinitely—since it is to their benefit—or having such a vague definition that it will practically be a ticket to continue their misconduct. By pivoting away from objectivism and turning to subjectivism instead, these concerns are groundless. As the ethical training is outsourced to the individuals, developers can no longer bluntly crawl their way in or lobby their way out. An unfair advantage that might or might not be used will no longer be a temptation for them, as it will be impossible to forecast and define parameters for every different individual personalized algorithm.

Etzioni and Etzioni argued that there are two ways society can reach to a decision: either collectively or individually (Etzioni & Etzioni, 2017). An example of a collective decision is the road instructions that everybody adheres to in order to drive safely from place to place. An example of an individual decision is, by using the same driving context if a driver should stop to get a hitchhiker from the street or pass by without helping. Putting this into the Artificial Intelligence ethics context, collective decisions are, as previously argued the objectively defined moral values that are few and vague. But Etzioni and Etzioni, to their credit, in order to argue for the individual decisions, proposed the personalization of AI agents as a solution, or AI algorithms that learn from their interactions with the users and adjust accordingly (Etzioni & Etzioni, 2017). One might argue that this is the way that AI agents work at the moment, much like Netflix’s recommendation algorithms. While this is a good first step
towards Etzioni’s & Etzioni’s model, this idea has to be both further developed and applied in different circumstances, like autonomous cars, for example.

**Personalization Implementation**

While this is very close to what I am proposing, there is an important difference. I find it problematic for users to simply personalize the algorithms, as a lot of users will not be tech-savvy enough to do so, even if the personalization is user-friendly. Given a certain amount of choices, on say, how would one react in a hypothetical scenario might be easy to answer but it does not provide the richness that I have in mind, coming from Protagoras’ thinking. Developers can neither forecast something that has not happened nor give alternative solutions for every possible scenario. Thus, a danger appears with this approach; the danger that these agents will not be sophisticated enough to be trusted. I also find it problematic for machines to simply learn from their interactions with the users as this will allow for an immense amount of data collection, a problem that has been apparent in contemporary literature, and might be even bigger in the future, if practices like these are implemented. The danger of extremely detailed profiling of users that might don’t want to be so lures, as well as a huge potential opportunity for developers to monetize this aspect and become even more powerful.

My proposal combines both of the two aspects, in a way, but tries to get away from the potential problems as well. The personalization of each user will be done in two ways. The first one will be for the algorithm to give its recommended solutions to a number of different scenarios, and then the user evaluating if this is something that she would have done herself. In this way, the algorithm can collect enough data that will be able to draw conclusions as to how to react in real life scenarios. Certainly, data collection will still be a problem, but instead of Etzioni’s and Etzioni’s proposition of being trained to real time human interactions, this way the environment will be much more sterile for developers to collect other kinds of irrelevant to the case data. The second one will be to put the already trained AI agent into a real-life simulator with its human user with it. There, minor tweaks on the algorithm can be made by correcting the algorithm when wrong and encouraging it when right. In this way, both data are much safer, and the AI agents are sufficiently personalized and ready for real-life. I would recommend that an optional choice of if the human agents want their AI agents to collect data from real-life interactions be present, lest one wants to constantly give feedback to their AI agents to be on the safer side of things, at least for the compatibility of their AI agent with their morality.

Another important aspect that needs to be considered is the fact that some users might be able but unwilling to do so. Here, transparency is key. Much to most people’s disagreement, AI agents are viewed as black boxes. Not wanting to willingly give away so much data is a fair concern. Developers here need to look at the bigger picture. They need to convince their audiences that their worries hold no ground. Being constantly open about their projects, their uses, and the data can seem like giving away trade secrets, but there is an
important difference; in this situation, instead of losing market value, they will actually gain public trust and therefore new customers. In contrary to business models where zero-sum games are seen as the only approach, by initiating a new player in the game -consumers- this situation turns to win-win. Companies will gain their much desired market penetration and consumers will equally gain their much desired knowledge of how their data are treated.

Potential Critique to the model

It is apparent that this is a proposition relying heavily on subjectivism, and thus, critical arguments against it are relevant to this model as well. An important one is that by having no universal laws, individualism becomes far too autonomous, to the point of dangerous. It would allow, in a sense, advocates of hate speech, terrorists, segregationists and other extremists to be assisted in their campaigns and quests by algorithms that will share their hateful ideologies. While this is an honest concern, there are two arguments that -to my mind at least- manage to equal the scale. The first one is that, as explained earlier, some few and universally acclaimed objective rules will be put into these algorithms; the value of human life will be over and above any kind of personalization and/or tweaking towards a certain mindset. While this is a clear deontological viewpoint, as I argued beforehand, I do not wish to completely contradict normative ideologies already discussed by other authors but rather take their best arguments and incorporate them into my model. In this way, part of the problem is effectively resolved. But surely, there will be multiple cases where human lives will not be put directly in danger but will either be indirectly so or threatened at different aspects, for example with marginalization. In those cases, it helps to think of it another way. Had there been brainwashed humans instead of algorithms helping out those extremists, there would have been no difference, save the -arguably- greater effectiveness of the AI agents. In other words, worldly operations will not change for the worse, simply because those actions would have been -and still are- possible without the introduction of Artificial Intelligence. Algorithms are going to be used as a helping hand, not as moral entities.

In any case, these problems are much more extreme than what one might initially think. In Hans Rosling’s book “Factfulness” there are quite a few examples where generalizing, relying on averages, or even taking extreme cases as the norm is quite problematic (Rosling, 2018). For most of the extreme-case scenarios that one might think about on how subjectivism can hurt humanity, there will be many more examples of people using it for good. Similar discussions have been held before about the use and misuse of the internet (Howard & Jones, 2004), video games (Bavelier, et al., 2011), television (Fiske, 1986), and radio (Mollgaard, 2012). It is obvious that this discussion follows every new technology and

---

9 I have to remind here that this project concerns ANI (Artificial Narrow Intelligence), therefore questions of AI acting and making decisions on their own are circumcised. AI agents will only be able to excel in one task.
the way it “alters” the world, or how malicious agents can create havoc among civic societies by effectively gaming these technologies. While both statements are based on true premises, the fact remains that by learning from past mistakes, every new technology transforms the world into a better place. It is also equally true that every new technology keeps getting harder and harder to regulate and put into normative standards. Thus, this essay proposes a model that will require minimum regulation and maximum autonomy, without wreaking havoc because of an important argument, next to be discussed.

Subjectivity does not mean unlawfulness

Having a Protagorean subjective morality does not mean that AI agents will not abide by certain laws and regulations. Laws that will penalize not only them but their human agents as well\(^\text{10}\), since they will be responsible for its upbringing. In this way, the problem of bouncing the problem between different parties involved -developers, users, agents- will be terminated, as will be much of the desire for people to do harm without consequences. By acquiring an AI agent that can be put to malicious use and not involve its human user into some form of penalty might be a recipe for disaster. On the contrary, by having the human agent solely responsible for its AI agent’s actions, much more thought will be put both when personalizing the algorithms and how to use them. To my mind, it makes perfect sense to hold human agents responsible since the AI agents will practically be their clones. If they don’t feel that they can trust them enough, maybe said algorithms are not yet ready and their deportation into society should not be rushed in the first place. On the other hand, a case can easily be appealed now, not by adhering to the artificial human emotions that would have been coded into AI agents but by adhering to the human emotions of the human user behind the algorithm.

Implementation in Autonomous Cars and more

To give a concrete example, I will refer to autonomous cars. I need to show an example of an AI agent which is complex enough, controversial enough, and not put in good use so far. The objective rules that I stated that it has to abide by are a) the highway code and b) United Nation’s “universal declaration of human rights” (United Nations, 1948). The personalization aspect will be -as explained above- in two stages. In the first, the AI agent will have a set of different scenarios where it will give a random answer. A question might be for example, “The road is open up ahead so I (the AI agent) reach the speed limit to make your travel time faster. Is this how you want me to proceed? Yes | No (please specify conditions)”. This way, the

\(^{10}\) This refers of course not to all AI being used at the moment, but only to the ones that report to humans. While it seems narrow in terms of application, ANI is being developed in all areas. Autonomous vehicles, personal assistants, and work automators are just a few examples to showcase the versatility of ANI.
algorithm gets a primary idea of what its human user wants to do in general circumstances. After that, a simulator running on any kind of monitor -where the human agent will be completely safe- tweaks the algorithm based on some imaginary but “unpredictable situations”. One of them, for example, could be the trolley problem which I analyzed previously. The AI agent will finish its action and then, the human agent can readjust it if she does not agree with how the algorithm handled the situation.

This is just a crude example, of course, but it can be put to use in various instances in technology we see evolving or available already. The teachers’ evaluators and AI judges could highly benefit from such a model because it will give a much more humane approach to a faceless machine. At the same time, cases will be able to be appealed since it will be accepted that AI agents are not omnipotent -as it mainly happens in objectivism- but prone to mistakes like their human agents. This way, the concern of omnipotence is eradicated. At the same time, if I take another example, with CV screening algorithms, having a personalized AI agent, there will be much greater diversity as to how do these algorithms scan, what are they looking for, and how they are evaluating results. Instead of a single program, running “objective” commands, therefore eliminating human peculiarity, agents that are much more humane, much smarter, and -depending on their user- much more kind-hearted, cynical, or open to alternative options.

A final example I would like to give derives from a crossing between ANI (Artificial Narrow Intelligence) and AGI (Artificial General Intelligence). The technology for autonomous robots helping people in need is already embedded into several institutions, even though several concerns arise as to how do these machines interact with “their patients”. In situations like these, for reasons of practicality, it would be impossible to personalize one robot per patient, as one AI agent handles several of them at the same time. Instead of developers consulting doctors or nurses on what are some features that these robots should have in order to do a better job, it would be far easier to personalize said robots to the doctors themselves. This way, patients can feel that they have an assistant that has a similar personality to their doctor, a person that they trust and see as an authority anyway. It would be much more soothing and humane to interact with something that has a certain morality; not one that is “canned” but rather one that is tailor-made after their actual human-being carer.

All in all, the model proposed based on Protagoras’ doctrine will be personalized after every unique user, thereby allowing for great variety of voices and viewpoints without marginalizing the weaker marketing groups or oppressing alternative minds. While this seems like as much work as training a human assistant there are two important differences: A) the algorithm will come bias-free. Even though humans come preconditioned with values, beliefs, and ideologies, AI agents are a true tabula rasa. This is important because users will not risk conflicts between their own ideologies and their assistants’. By the time the AI agents is trained, they would expect to be in complete alignment with their own philosophy. B) consistency is key. While each person is different (that’s why there are rigorous hiring
processes in the industry to find the right person for an opening), AI agents’ unique selling proposition is that they are all the same; a user knows exactly what she is purchasing. Much like shopping clothes from a certain brand, one expects consistency in quality, exactly as it has always been. Similarly, while humans are mostly unpredictable, difficult to judge, and impossible to duplicate, the same cannot be said for robots; they are always the same until a user interacts with them and shape them however she wants to.

This model will be built upon a few but universally acclaimed human rights, and in this way, it will allow for a solid legal framework to be established, since the human agents will be held solely and directly responsible for the misconducts of their AI agents. Again, this derives from normative ethics which I do not scope to undermine but use effectively when appropriate. Additionally, the model will bring to the surface arguments that, as Protagoras claimed with his wind example, are impossible to imagine without a reported experience, in order to adjust this whole discussion -if needed- in case the need arises. Finally, its use for nefarious purposes will be limited, both because human agents will be considered responsible and because zooming out to see the bigger picture, every technology so far has been used for -mainly- upgrading social standards. It would be irrational to think that Artificial Intelligence will not contribute to humanity’s evolution, as long as reins are there to keep it on track. Since ethics and morality when used for rules and regulations are very much like horse reins, allow me this last simile. Holding the reins too loose, the horse might decline from its course. Holding them too tight and it will stop. A good balance is needed in order to keep the horse steady but moving; a balance that is greatly achieved by Protagoras’ subjectivism, encapsulated and tailored to frame the Artificial Intelligence ethics debate. Before concluding, I would like to refer briefly to a second Protagorean fragment because I think it adds significantly to the discussion.

The two-logos fragment

Protagoras was considered among one of the sophists (Guthrie, 1969). Sophists were, among others, private teachers that could teach prosperous young boys important lessons, mainly rhetoric. Aristotle, as I have shown before, created the whole doctrine of Logic in order to analyze, master, and teach the art of rhetoric, as it was supposed to be the most useful skill one could have in the Athenian conventions. Much like Heraclitus before him, Protagoras argued in his infamous two-logos fragment (Schiappa, 2003) that he could make the weaker become stronger11 (Schiappa, 2003). This fragment has been severely misrepresented and Protagoras has been attacked both ad hominem, marking him an amoralist whose only concern was to win over arguments in oratories, and for his arguments, claiming that this contradicts with the objectivity of the world (Schiappa, 2003). Schiappa is, though, not

---

11 Author's translation of the original “τὸν ἄττω λόγον κρείττω ποιεῖν”
convinced. He claimed that Aristotelian critique is anachronistic, as these words did not mean the same thing in Protagoras’ time (2003). What they really meant was weak and strong (as I have loosely translated them) and not “good” and “bad” as the same words meant in Aristotle’s time (Schiappa, 2003). What Schiappa really showed is that Protagoras, instead of being an unethical instructor whose job was to teach bad people how to promote their ideas, he was one who could give an equal saying to an unpopular belief, independently of how right or wrong it was.

If I trace this discussion back to Artificial Intelligence ethics, Protagoras’ standpoint is clear; all voices have to be heard, independently of their popularity. My model promotes this notion ideally, as it creates space for autonomy while takes precautions that the right to be heard is not abused. Indeed, autonomy is created by allowing for personalization of AI agents, while precautions are taken through putting responsibility to human agents for their algorithms’ actions. Noble’ and Benjamin’s rightful concerns about the oppressive nature of Artificial Intelligence are being addressed thus clearly (Noble, 2018; Benjamin, 2019). Important is also the distinction between each voice having a saying, and each voice having an equal saying. It would be impossible, even dangerous to give an equal saying to every different ethical standpoint. However, no one can feel marginalized since their personal AI agents will technically be 100% compatible with their ideology. Sure enough, when people interact with other people’s algorithms, they might feel differently, but again, this is the way human societies have been operating so far; different people exchanging ideas, interacting, and possibly disagreeing.

Since thus these AI agents will be personal robots in essence, an interesting question arises; would it be realistic to believe that people would want to spend the time it would take to train such a robot? I will try to answer this question using all four philosophical schools of thought used in this project to address many aspects of the problem. From a deontological point of view, people would do this not only for their own sake but because this would be the “right” thing to do. In a society already infiltrated by AI it would be our moral obligation to teach these AI agents how to behave ethically. From a consequentialist point of view, if this idea’s benefits outnumbered its costs, it would have to be chosen. So far I have argued that the benefits are indeed greater. From the virtue ethics’ point of view, since it benefits both the individual and society while it paves a way for a middle ground between two extremes, it would be equally compatible. Finally, from a subjectivist’s point of view it would make sense, since it is one of only a few models so far that argue in favor of relativism. All in all, philosophically speaking, people should not object to this model. Of course, I do not imply that this answer proves that people will actually accept this model. A holistic answer like this would require a second project this size. My answer partly addresses the philosophical aspect of it, since this is this project’s focus.
Conclusion

Summary

The initial idea about this project was split into four parts. The first was to collect as many voices as possible from scholars arguing about Artificial Intelligence ethics without making a distinction on the critical or supportive standpoints of each author. The second was to map them out in different ethical schools of thought in order to have a clearer delineation of the influences and fundamentals of each researcher. The third was to try and connect all ethical school of thought to the major three classical ancient Greek philosophers, Socrates, his student Plato, and the latter’s student Aristotle. The final part tried to initiate Protagoras into the discussion, as I believe he is relevant to the discussion. A crude solution has been offered to the ongoing debate even though it is still far from perfect.

I wanted to bring as many sources as possible - due to time and space constraints - because a debate is as good as the diversity of its arguments. It would be a grave error to cite only the biggest names and ideas as I would potentially miss out on some authors which, even though have excellent contributions to the discussion, have not yet received much attention (see Etzioni and Etzioni, 2017). Bringing all voices together would be unrealistic expectations since the literature is not only vast but continuously updated. When one chooses a very popular societal and scientific issue, finding enough material is hardly a problem. However, selecting the right amount and quality is a very demanding issue. I do not declare that I covered the subject 100% in Literature Review as that would take several projects this size; rather, I commented on selected bibliography depending on the variety of arguments, ease to retrieve, and of course access, since as a student it is limited.

For the three main schools of thought that I used as umbrella terms, namely Deontology, Consequentialism, and Aristotelian Virtue Ethics, generalizations had to be made, and I am aware that the small definitions I provided them with are a bit narrow and simplistic. It would be, however, out of the scope of this essay to have an extended summary of each school of thought, as they have only been used to demarcate the Artificial Intelligence arguments analyzed in the Literature Review, and not being analyzed and evaluated in themselves. In other words, even though I did comment quite extensively on the values and vices of each school’s ideas, I did so only in relation to Artificial Intelligence, having as my sole goal how to solve this ethical riddle. The debate among philosophers as to which school is more appropriate to society in general has not yet been resolved and it is not a part of this essay. It follows then, that my critique on each school focus only on AI ethics and should not be taken out of context as a general critique of the whole doctrine.

The connection with the ancient Greek philosophers is something that, to the best of my knowledge, has not yet been argued by any scholar so far, at least in total. Surely enough, separate connections have been made, like for example Scarre’s connection of Utilitarianism
and the ideas of Aristotle and Epicurus (Scarre, 1996), but to argue that all three major school of thoughts have derived from the three ancient Greek classical philosophers is something that I could not find. I hope that I sufficiently argued about that in the section Ancient Greek Philosophers because, as I could not stand on someone else’s work, I had to dedicate enough time and space. After all, this is one of the most important claims of this paper, as without it, the transition to Protagoras’ doctrines would seem artificial and out of place.

Similarly, Protagoras’ arguments have been marginalized, as both society and scholars consider the sophists to be irrelevant to ethical discussions. Sophists have been misinterpreted so far, mostly because our main sources of that time derive from Plato and Aristotle which were against the sophistic movement (Schiappa, 2003). Both of them misinterpreted him either intentionally or unintentionally (Schiappa, 2003). An example would be my previous argument about Aristotle’s definition of good and bad in Protagoras’ “two-logos fragment” as something that has changed meaning over the course of time. Since it meant something different to Protagoras’ time, Aristotle misinterpreted him; as a result, since Aristotle is a genuine primary source while Protagoras’ work is known through others’ texts, Aristotle’s definition received more attention.

Schiappa informed us that Protagoras was active in the Periclean democracy, as he was the lawmaker for the Athenian colony Thurii (Schiappa, 2003). He is also believed to be one of the "giants", whose shoulders upon the Athenian democracy was built (Schiappa, 2003). Protagoras believed that "areti" (or virtue in Aristotelian terms) could be taught, something quite revolutionary at the time (Schiappa, 2003). Before the 5th C. BCE Athenian democracy, the oligarchs were seen as having innate “areti” when they were born. Protagoras was the first to democratize this notion, by initiating his teaching method which relied both on instruction and punishment (Schiappa, 2003). The role of the teacher was to instruct, while the role of the law was to punish. A punishment, not for revenge, but for rehabilitation for both the aggregator and the spectators (Schiappa, 2003). His primary example for the teachability of “areti” was the Greek language, as it is a virtue being passed to the new generation by same means (Schiappa, 2003). For the ancient Greeks, both oral and written language was a virtue of high importance. Since babies cannot speak or write the Greek language, they learn to do so as they grow up; hence, “areti” is something that both can be taught and is continuously being so without people realizing it.

**Limitations**

This model, as most models do, has some limitations that should be taken into account when it comes to both design subjectivism into Artificial Intelligence agents as well as implementing them into society. It should not be forgotten that personalized algorithms do exist at the moment, even though they are in a different format than the one I am proposing. Google maps is one such implementation, as it not only allows for users to input preferred data, *home address, work address* and many more, but it also allows for *favorites*, recommending and giving feedback to *recently visited places*, as well as choosing the best
route depending one’s usage; a biker will get a different road than a car driver or a runner, or even two drivers will get a different recommended route depending on their previous choices of paying tolls or avoiding the city center or highways. Nonetheless, even though commercial applications of personalized AI agents exist today, they are on a much lesser scale and magnitude than the algorithms I am interested in. All in all, the difference between a recommended route and an autonomous car is significant. Instead of proposing trivial information, in the second scenario AI agents will be active members of society, at least in this context.

It is apparent, thus, that even a subjective model is not fit for every agent. Surely enough, with enough tweaking and attention to detail a similar model might be able to help in more ethical considerations using AI agents into society, but for the most part, the Protagorean solution concerns mostly ANI agents that hold important societal positions. By the time humanity achieves AGI, if it ever comes to fruition, this model might no longer be valid. Therefore, it is important to not be taken out of context, as this is quite a radical proposition that is very easy to face opposition. While in some cases it will be rightfully so, I think it is more important to focus on where it can and cannot be applied, and not assume a priori that it is a one-size-fit-all solution. That would be an oversimplification, and even if possible, well beyond the scope of this project.

On a similar level, the fact that extreme cases, where people will potentially use this subjective model to find a hole in the system and use it for nefarious purposes might be small, it does not mean that it is impossible or insignificant. Rightfully so, these cases must be well thought in advance -if possible- or reported extensively so that the proposed model can be altered if need be or readjusted for the sake of safety. It has to be pointed out though that no system is flawless, even if objective cases are taken into account, like programming -which uses Logic- or banking systems -which use statistics and mathematics. Several major breaches have been reported for the former, as well as multiple misconducts on the latter. A system made from humans will always be prone to error. It is my viewpoint, though, that by having the best alternative of an open-source program where everyone will source in data and give feedback, the chance for overlooking details or trivial mistakes that might prove disastrous is minimized. Malicious agents will exist, no doubt, and this model might give rise to some novel ways of misbehavior but it seems to me that this model gives the highest percentage of success with the minimum amount of effort, while, at the same time, promoting equality and diversity, much like Protagoras did in his time.

It has to be noted here that, even though Protagoras democratized knowledge, societies, and minds alike, he was getting paid quite extensively for his services. As a consequence, not everybody had the opportunity to be instructed by the great teacher. In most cases, the rich and powerful were the ones that could afford his services. The result was that since most people did not have the luxury of education, the few and well-instructed had now even greater power than before. However, a general societal trend has been observed, where by
the time people knew an instructor existed, one that could potentially teach them how to be more active citizens and lead more prosperous lives, decided to do everything possible to get this knowledge, resulting in the Periclean democracy that western societies are built upon.

As in ancient Athens, so in contemporary societies a subjective ethical model on Artificial Intelligence may claim that will make sure that different voices will be heard, but the economic barrier of who might own such agents is still relevant. However, societal changes need but a spark, one that will give hope to the oppressed and marginalized, one that promises equality even if it is at a great cost. As did the Athenians, so will citizens of the 21st C. find ways to take a fairer system over a lesser one and incorporating it into society. Ways that I am unable to imagine as I am a single entity. The imagination of an extensive amount of people cannot and should not be underestimated; brainstorming takes numbers to be most effective. It has been apparent with the usage of previous technologies that managed to connect people together, like the internet. The internet is a good example in this discussion as it is, arguably, another subjective, Protagorean concept in terms of its ethics; and even though it does not come without problems, the impact it had in democratizing societies is unprecedented. I forecast something similar for AI agents if a subjective ethical standpoint is implemented.

**Final notes**

This project is only a theoretical contribution and consequently, further research is demanded to test the implementation of the proposed model. Views of the experts on the field upon this model should be documented and analyzed in order to have a greater variety of arguments. Quantitative and qualitative studies on the perception of the public could be a way forward as well, in order to test if citizens are as optimistic as I am to personalize AI agents. Their opinion on if they are willing to have Artificial Intelligent clones of themselves is of critical importance, as without their consent this idea will be fruitless. Finally, governments and opinion leaders should have their own voices heard, because they might see something that academics cannot see or point out from their experience strong and weak points that need to be reconsidered. Whatever the case may be, as argued above, the greater the number of people commenting on the model, the better it can serve society. Different opinions will of course be voiced, some of them with malicious intent, but the whole essay has been focused on subjectivity all along. It would be hypocritical to muzzle critics, independently of their reasons for being critical.

Artificial Intelligence is something that should unify humanity, something that should drive progress. Instead of reminding all of the human vices that resurface every time a new technology comes into play, Artificial Intelligence should unite people; all in all, if the technological optimists’ forecasting is true, humans might not be the only Intelligent species on the planet for long. That alone is a fact that should remind people of their mortality, of the -arguably- much bigger things that unite them instead of those that segregate them. In the end, every different way humans devised to separate themselves into clusters are much
lesser, much more powerless than those that bind them together. Malicious agents will always exist, but this should not be a distraction from the fact that most people want to contribute to humanity’s upward motion. As Mahatma Gandhi said it best, “You must not lose faith in humanity. Humanity is an ocean; if a few drops of the ocean are dirty, the ocean does not become dirty.” (Gandhi, n.d.) Let Protagorean relativism be an ethical guide in this discussion in order to move forward.
Bibliography


Gunkel, D. J. (2017). The other question: can and should robots have rights? Ethics and Information Technology, 20, 87-99.


