Killer Apps and Killer Robots

An Ethical Framework for Answering Questions of Automation

Peter McKay, Nisansa de Silva, Elizabeth Fuller, Bradley Green, Gautam Sondur
Information and Computer science Department
University of Oregon
CIS 590 - Winter 2016

I. Introduction

"Sweating as he turned here and there to his bellows busily, since he was working on twenty tripods which were to stand against the wall of his strong-founded dwelling. And he had set golden wheels underneath the base of each one so that of their own motion they could wheel into the immortal gathering, and return to his house: a wonder to look at. These were so far finished, but the elaborate ear handles were not yet on. He was forging these, and beating the chains out"

writes Homer in Iliad [1]. In fact, Iliad is not alone in ancient literature to describe such contraptions; mentions of automatons can be seen throughout other Greek texts [2]. Ancient Liezi text of China talks about an android singer presented to the fifth king of the Chinese Zhou Dynasty, King Mu of Zhou [3]. The 11th/12th century Indian text, Lokapannatti, talks about robot guards and soldiers [4]. Even though one might question whether these are about actual incidents or not, these records make one thing certain: the idea of self-working machines is not a new concept to humanity. It is, rather, an extension of the unquenchable human desire to make their lives easier; the desire that gave birth to the rudimentary tools that propelled us up the evolutionary tree, branching up and away from our animal brethren.

With the digital revolution spearheaded by the discovery of the transistor, we are now at a time when the very same stories, which one might brush off as fantasies when read in the above ancient texts, are slowly but surely becoming commonplace. In place of the auto moving tripods of Hephaistos in Iliad, we now have self-driving cars created by Google [5] and Autopilot mode enabled cars by Tesla Motors [6]. Instead of stopping at the creation of singing automatons such as the one from the Liezi text, we have created automated systems that can compose music [7] and create works of art [8]. We are using drones and other unmanned vehicles to wage war. Fully automated security robots are emerging [9]. In light of these developments, it is reasonable to say that we are going far and beyond the wildest imaginations of our ancestors.

The discussion about artificial intelligence has grown to the level of debating whether a future artificial intelligence that is comparable to a human in all walks of life (so that unless specified a human will not be able to distinguish between the said AI and another human [10]), is sapient or not. If we do agree that the AIs are sapient, from that agreement raises the question, "Is humanity using slave labor when it employees sapient robots to do tasks for them?" If a robot commits an offence that is punishable by law, who is liable? Is it the creators of the AI program or is a sapient AI able to take legal responsibility of its own actions? If it is the latter, what kind of rights would the AI have in a legal setting? If it is the former how is it different from the "Sins of the son" fallacy where parents are held liable for crimes committed by their children.

Like every scientific breakthrough before it, automation has created a dichotomy; one group consisting of optimistic scientists such as sir Arthur C. Clarke who claim "The only way of discovering the limits of the possible is to venture a little way past them into the impossible" [11], and the other consisting of scientists such as Stephen Hawking, Elon Musk, and Bill Gates who are rather pessimistic about the advent of artificial intelligence and thus warns that humanity will have to be diligent as the

technology develops [12]. Those who are not optimistic about the emergence of the automation also claim that it need not be developed to yield killer robots or artificial general intelligence [13] for automation to have a deep social impact. They clame that even a small amount of automation might have far reaching consequences. The objective of this paper is to analyze that social impact in an ethical standpoint and propose an ethical framework that can be used to mitigate the possible undesirable consequences of automation.

In section II we will discuss the socio-ethical impact that the industrial revolution had on the society given that fact that it is the only adequately documented closest approximate technological automation based transformative period of human history that we can compare this new contemporary transformation with smarter and almost autonomous machines. Then we discuss the situation that the society and technology currently stands in section III. Moving forward, in section IV, we analyze several modern approaches as advanced by a number of prominent individuals and groups: Singularitarians, Bill Joy, Jason Lanier, the Amish, and Deep Green Resistance. For each of these proposed solutions we discuss the analysis that has led them to such a solution, and the advantages and shortcomings which characterize those approaches.

In section V we will discuss the solution that we propose in this paper. The solution is explained in terms of; Identifying Stakeholders, General Classes of Stakeholders, and Defining a Coherent Value System. We explain how our proposed framework can be used by way of the auto-driving car use case in section VI. Finally, the section VII gives the conclusions of the paper.

II. Background

Later on in our paper, we will put forth an ethical framework for use in answering questions pertaining to automation and related technologies. As an important part of that framework, we aim to analyze who may be affected by changes in the state of the art of technology, and how those effects may ripple out to change society at large. While these analyses would be much simpler were it not the case, the fact of the matter is that technological innovation is part and parcel of human society. Advances in technological innovation cannot be extricated from the social conditions which gave rise to such innovation, nor can they be separated from the social consequences which result from the adoption of that technology. If we wish to base our analyses in historical fact, it seems prudent to first review some basic history with an eye towards the factors we intend to consider. If we wish to see a preview of the impacts that automation will have on the workforce, and consequently the society at large, it seems prudent to look at the ways automation shaped society as its effects on the economy were first made manifest.

Let us consider, for a moment, the steam hammer. The steam hammer, a powered hammer for forging and stamping of metal parts, was invented in the mid-1800s as increasingly large and complex machines necessitated the fabrication of larger and heavier components. It made routine the construction of many works that, only years before, would have been considered marvels of engineering and concerted human effort. It contributed to the construction of great ships, the production of delicate clockwork, and a massive increase in the industrial productivity of Western Europe that saw an increase in the availability of cheaply manufactured goods formerly restricted to only the richest members of society. It also contributed [14] to a sharp decrease in the leverage that skilled engineers held over their employers, resulting in drastic consequences for working environments, and sharp increases in child-labor. James Nasmyth, one of those credited with the invention of the steam hammer, made the following statement regarding the power of automation in the modern (as of 1851) factory:

> "The characteristic feature of our modern mechanical improvements is the introduction of self-acting machinery. What every mechanical workman has now to do, and what every boy can do, is not to work himself, but to superintend the beautiful labour of the machine. The whole class of workmen that depend exclusively upon their skill is now done away with. Formerly, I employed four boys to every machine. Thanks to these new mechanical combinations, I have reduced the number of grown-up men from 1,500 to 759. The result was a considerable increase in my profits." [15]

From the specific, we proceed to the general. The example of the steam hammer (in particular, the contrast between its positive and negative social effects) raises one of the first difficulties we will be forced to surmount in the course of developing our ethical framework. In the case of the industrialization of the workforce of Western Europe, those who wished to deploy greater degrees of automation in their factories were directly at odds with those who work in those factories. This is reflective of the effects of automation, not just on the market for the product they create, but also on the labor market. Labor, like any other commodity, has a price, which fluctuates based on supply and demand. When the owners of the factory can produce just as much value from less work, the demand for that labor goes down, and its price drops. Those workers have less leverage, and, as observed by Gaskell, their working conditions suffer.

When balancing the gains in technology and standards of living that have resulted from this march of progress, we see that there are two sides of this coin we must consider. The society as a whole may benefit, but what of the workers that toil in the factory? Does the gain outweigh the harm? Fortunately, in this particular case, we are not forced to make such a decision: social and cultural pressures intervened. The labor movement, in an internationally coordinated effort, slowly but surely pushed many governments into enacting legislation that protected workers from unfair practices on the part of their employers.

And yet, we are faced with a somewhat tougher questions, when we step back and consider the broader context around the industrialization of Western Europe. Industrialization is a somewhat titanic effort, it requires the investment of massive amount of capital to begin to see the advantages resulting from leveraging the economies of scale that make automation profitable. Different societies have achieved such a concentration of wealth in different ways. China, in the course of The Great Leap Forward, began industrialization on a massive scale by nationalizing all production, taking central command of labor and distributing it according to a grand plan for industrialization of the workforce as a whole. The Soviet Union took a similar approach, nationalizing industry and placing it under central command, but then decentralized the industry geographically, exporting raw materials to nearby states as it absorbed them, thus creating a demand for the growth of industry in those formerly independent states that drove their growth further outward until they started bumping up against NATO countries. Western European countries accumulated wealth in privately held trading companies by establishing an early foothold in mercantile capitalism through the Triangle Trade. That is, they establishing trade routes between Europe, Africa, and the Americas, alternating in the shipment of slaves, raw materials (such as sugar cane grown by slaves), and manufactured goods (such as rum produced from last year's sugar cane).

In each of these situations, there exists a contradiction between the forces acquiring enough wealth to assemble an industrial base, and the people that had to give up that wealth. In China and the Soviet Union, the costs were paid by the citizens of those countries, and they reaped the benefits slowly as their standards of living increased fairly evenly. In Europe, the standards of living and availability of luxury goods skyrocketed much more rapidly: after all, the cost of automaton was offloaded to somewhere else: Africa and the American colonies. In the Soviet Union and China, the analysis looks simple: did the country gain more than it lost? China suffered a number of famines as the Great Leap Forward encountered organizational difficulties, but over a period of 30 years, the average lifespan of a Chinese citizen nearly doubled.

When, however, we look at a situation like British industrialization, we now find that we must take a look at a much more diverse selection of groups in analyzing the impact of technology on the society. As globalization has blurred the lines between nations, so too has information technology led to a vastly more interconnected world than we have ever seen. A change in technology today can lead to nearly instantaneous changes through enormous swaths of society. In the immediately following sections, we will further examine the current

state of the world as it has evolved, the approaches that some groups take in analyzing that world, and the steps they deem necessary due to such an analysis.

III. The Current Scenario

On the 12^{th} of March 2016, history was made. Lee Sedol, a professional Go Player who some consider to be the best player of the last decade, lost a game of Go. His third consecutive loss in a best-of-five competition. His opponent was AlphaGo, an Artificial Intelligence designed to play Go [16]. To anyone paying attention to this field, it does not seem terribly surprising. After all, IBM's Watson managed to outsmart previous human jeopardy champions, [17], and it's been decades since the last time a human was able to compete witha chess-playing AI. But there is something that separates Go from these other games.

Firstly, a game of Go can represent a vast array of possible states, approximately 10^{768}, a number more suitable for counting atoms in the observable universe than counting Chess positions. This rules out brute-force techniques for solving the game. That means that we can't just solve Go, we need to teach the computer how to actually play Go. This leads to our second point. This is a significant milestone in the trend of using computers to replicate mental processes that were earlier considered to be the sole dominion of the living beings, particularly humans.

As mentioned earlier, we live in the Information Age, and changes have a tendency to propagate through our society fairly quickly. Globalization, and the subsequent transfers of technology enabled by the Internet, allows for a proliferation of ideas and techniques that would have been unheard of even a few decades ago. Automation's strongest argument has always been that it allows us to reduce manual labor and improve the quality of human lives by allowing humans to direct their labors towards the kind of work that assembly lines cannot replicate: matters of intellect. Some may argue that this is no longer the case. Many proponents of automation have pointed to the migration of people from unskilled and semi-skilled labor to more skilled labor which rely on mental processes that could not have been automated at the time. It is evident that this sacred boundary is moving, and, depending on whom you ask, may already be at risk of dissolving entirely.

The advent of applications that have improved accessibility to services, from planning your week to simply setting an alarm or a reminder to do your taxes, all these tasks have been automated already. The automation of these activities begs the question of whether this would make humans obsolete in performing these activities. The public sentiment does not seem to reflect this. In a study carried out by the *Pew Research Center* [18], although about two-thirds of Americans believe that computers and robots will be performing most of the work being done by humans in the next 50 years, nearly 80 % believe that their current jobs will survive the transition..

A likely answer to this is skewed figure is probably to do with the general outlook most people have on technology. In a 1996 *Washington Post/Kaiser Family Foundation/Harvard* poll, 70% of Americans said the increased use of technology in the workplace was good for the economy. In a 2010 *Allstate/National Journal* poll, 79% said that information technology was extremely or very important to creating economic growth in the U.S. They may not be entirely wrong, but it should be noted that the opinion of the general public is rarely aligned with the most ethical option.

Even though AI may seem to form only a small part of the problem we have so far discussed, it makes up the primary difference between the *First Industrial Revolution* and what some are calling the beginning of The Second Machine age, as more and more mental processes can be performed by computers with greater speed and correctness. Technology Heavyweights like Elon Musk and Stephen Hawking have openly spoken about the perils of the technology when it is used for warfare [19], expressing concern about the possibility of human extinction if left unchecked. The extreme views on this issue are well documented, as we will further elaborate in the next section, but our aim remains to balance these views to ensure

a smooth transition. No situations is without its winners and losers, so the best we can do is ensure that the current march towards an increasingly automated society is done in the most ethical fashion. We wish to move forward without favoring the interests of some people over others, and ensuring that we emerge with a holistic view of the situation.

IV. A Review of Several Modern Positions

A. The Hope of the Singularity

The notion of the singularity was first popularized by computer scientist Ray Kurzweil. The singularity is a future time period when technological change will occur so rapidly and it's impact will be so large that human life will be irrevocably changed to such a degree that we will be unable to make any predictions about the world beyond the singularity. Whether the singularity is dystopian or utopian depends really on an individuals point of view.

The underlying insight behind the singularity is that the rate of change of techonology continues to accelerate. Furthermore, the myriad ways of leveraging technology are expanding at an exponential rate. The singularity is when information-based technologies will include all human knowledge and proficiency, including pattern-recognition powers, problem solving abilities, and emotional and moral intelligence.

Ray Kurzweil describes the singularity as that which

> "represents the culmination of the merger of our biologicacl thinking and existence with our technology, resulting in a world that is still human, but transcends our biological roots. There will be no distinction between human and machine or between physical and virtual reality." [20].

The claim of singulatarians is that we are in the beginning of the transition into the singularity at the knee of the curve. Perhaps the most alluring aspect of the hope of Singularity to many is the potential to conquer death itself. They hope to reverse the process of aging using biotechnology, and eventually free ourselves from all forms of death by uploading our minds into durable machine environments. Singulatarians recognize the risks inherent in future technologies such as genetics and nanotechnologies, but their approach to new technologies such as AI tends to emphasize the benefits over the risks. In their view, scientific and technological progress is seen as not just inevitable, but desirable no matter the possible cost.

To a singularitarian, they view their cause as nothing less than the most noble pursuit in history: the end of all human suffering. As such, simple problems like poverty or hunger are merely edge cases of our existing march towards their goal. So long as it allows us to bootstrap ourselves into the singularity faster, what care have we for those that fall along the path, when weighed against the whole of the future stretching out before us? The singularitarians represent a particular extreme with regard to their attitude towards technology. They acknowledge that technology may cause social problems, but claim that future technology will obviate those problems and all others.

B. Bill Joy

In recent years a new movement has begun to form in our society, one that most closely resembles the historical Luddite movement, but is different in some important respects. This movement is known as tech-skepticism, or neoluddism. Perhaps the most well known voice in this movement is Bill Joy, co-founder of Sun Microsystems. In 2000, he wrote an article that that reverberated through the tech industry, entitled "Why the Future Does Not Need Us" [21]. In this article, he articulated views held by many others that might have described themselves as tech skeptics. While he admited that there are fantastic benefits to technology, he made the point that there are also astronomical risks involved, risks we cannot afford to downplay.

In particular, he highlights that new technologies have a risk factor that previous technologies did not: robots, engineered organisms, and nanomachines all have the potential to self-replicate. Joy views this as an extremely dangerous threat, comparing the advent of genetics, nanotechnology, and robots (GNR) to the to the development of weapons of mass destruction of the previous era. Joy points out that the WMDs of the 20th century were powerful threats, but were ultimately self-limiting. Nuclear weapons, Joy noted, required access to highly

regulated raw materials, and shares requirements with biological and chemical weapon development programs for expensive large-scale labs and facilities.

At the heart of Bill Joy's argument is that the real difference between 20th and 21st century technologies is that these technologies are within the grasph of individuals or small groups. The new paradigm allows what he terms "knowledge enabled mass destruction," or KMD, which is only amplified by the ability of these technologies to self replicate. Joy speaks of the threat of extinction that these new technologies may bring, estimated by at something exceeding 30% [21]. Joy's answer to this dilemma between the fantastic benefits of technology and the astronomical dangers of the same technology diverges from the view that the proponents of singularity take. Bill Joy advocates for determining the course of the human species by looking at our collective values, ethics, and morals to determine which areas of knowledge are too dangerous to be allowed to be pursued and limiting the development of technologies that are too dangerous. Essentially, Bill Joy rejects the views of proponents of the singlarity that the existential risks are worth the possible benefits of pursuing knowledge in key areas. Furthermore, he warns that for technologies once we have developed these technologies, their self-replicating nature will make it impossible to get rid of them. According to the Joy and the neo-luddites, we are dangerously close to opening Pandora's box, and there's no sure answer as to whether it has any hope at the bottom.

C. Jason Lanier

While tech skeptics like Bill Joy are concerned with the collision between human nature and the capabilities of technologies like genetics, robotics, and nanotechnology, another movement of tech skepticism is critiquing technology as it is being used right now. One of the more prominent adherents of this movement is Jason Lanier, author of the best selling book "Who Owns The Future?" [22]. Jason lanier critiques current and future technology, not as bearers of existential risks, but as tools that allow their owners to usurp most of technology's benefits.

He questions how economic value is determined in the information age, arguing that it is information, not manual labor, that is the most valuable in today's economy. And yet, he opines, it is precisely this information that we are expected to give away for free. For example, companies like Facebook and Google, whose entire profit structure is based on using the information that they gain from the individuals that make up their customer base. He sees this system as providing disproportionate economic power to the companies that are able to access the most information on people the fastest. These companies take interests, demographic predilections, buying habits, and cybermovements all in exchange for only "free" admission into the social media networks that have become so popular that opting out can be detrimental to one's career. Instead of the customer, people are the product. He believes that people should be paid for the information that they contribute, that the vast profits companies make off of our combine information should be shared with the users that make it possible. To Jason Lanier, technology is not evil, and humans aren't evil. We have just found ourselves in a system that does not serve our best interests, and Lanier argues that we need to enact comprehensive reform in order to do it.

D. The Amish

The Amish in most popular depictions are painted as rural farmers. They are often protrayed as a culture who fear technology. However, this is a gross simplification of the relationship the Amish have with technology. Previously we have seen Bill Joy who is worried about the juxtaposition of human nature, underestimating the probability of catestrophic effects, and technology and Jason Lanier who critiques what he sees as the rampant corporatism with technology. In contrast, the Amish are not so worried about negative effects like pollution, but the effect of new technologies on their culture. The Amish view of technology is that technology and society or culture are created together and in the words of Sheila Jasanoff "both embeds and is embedded in social practices,identities, norms,

conventions, discourses, instruments, and institutions " [23]. The Amish regard technology as a force that has the power to reinforce social norms, control how people interact, or enable new forms of interaction between people. Where Bill Joy implore us to accurately weigh the net harm including the likelihood of catestrophic events with respect to technology and Jason Lanier insists that we need to insure that corporations do not gain most of the benifit of new technologies instead the Amish insist that any new technology is in line with their culture as well as reinforces their tradition, religion, and culture regardless of any objective potential benefits or harm associated with that technology.

E. Deep Green Resistance

Deep Green Resistance is a group primarily identified by their beliefs and behavior with regard to specifically environmental matters, rather than more general opinions about technology. Specifically, DGR subscribes to the notion that industrial civilization poses too great of a threat to life on the planet earth, both human and otherwise, to be allowed to survive. The ultimate goal of DGR is the destruction of industrial civilization and a return to an earlier stage of societal and economic development.

DGR's philosophy draws from the Deep Ecology movement, which holds that an anthropocentric analysis of ecology, defining other forms of life in terms of their utility to humankind, does not do an adequate job of describing the complexity of ecological systems. DGR takes this a step further, in declaring all life to be equal to human life. Starting at the problem of inequality among life, they turn to ideals similar to the political stance of anarcho-primitivism: the idea that human society, industrial capitalism in particular, cannot be reformed into a more beneficial form.

Such a philosophy stems from an agreement with Friedrech Engal's seminal anthropological work on early civilizations [24], specifically the assertion that early gatherer-hunter societies were unable to produce structural inequality without the ability to acquire surplus wealth. Unlike Engals, modern Marxists, and even their fellow anarchists, DGR and anarcho-primitivists do not believe that a society can ever reach a level of egalitarianism and freedom from oppression that they are comfortable with. Rather, they believes that the only truly equal societies that have ever existed were only able to function without the trappings of civilization. Therefore, they are willing to advocate any means to right the wrongs they see with the world. To them, technology is not evil per se. Rather, technology is another product of a system with which they cannot make peace, and as such, it must fall by the wayside.

Later in our paper, we advance an ethical framework for the analysis of decisions regarding technology and automation. In so doing, we emphasize that we must look to the past when examining possible consequences of an action, basing, wherever we can, our reasoning on historical fact. In this case, we would be remiss if we did not qualify our description of the views of DGR with the caveat that their views ultimately reduce to something chillingly familiar. A return to pre-industrial levels of technology would necessitate either the death of billions, or the restriction of new births to a fraction of their current levels. Without modern medical technology, it becomes a disturbing fact that disabled people will not be able to survive in such a world: this means DGR endorses either genocide or eugenics.

V. A Coherent Ethical Framework

A. Identifying Stakeholders

Ethical Analysis of an action, circumstance or rule must always begin with Stakeholder Analysis. Stakeholders are those who are affected as consequence of the action, circumstance or rule. In other words, the parties concerned have a stake in the eventual outcome of performing the action or implementing the rule. In our framework, we consider a rough categorization of factors that play into our stakeholder identification. Then, based on historical precedent, we observe the existence of larger classes of stakeholders. The interplay between these classes, in practice, follows historical trends that allow us to reason about larger societal effects without getting too bogged down in hypo-

theticals. That is, wherever we can, we wish to base our analysis not just on sound logic, but on historical fact.

The process of identifying Stakeholders is generally specific to the action and rule as well as to the circumstances that surround the process. Although our framework strives to espouse the Kantian Ideal of universality, it would be naive to assume that the stakeholders would be a constant across the board over time. To this end, in order to identify the stakeholders, we must try to obtain a holistic view of the situation, including our own biases and viewpoints in such an analysis. In order to do this, we identify three different types of factors that may affect the identification process.

- Cultural Bias:

  Cultural Bias has long hindered the adoption of technology. Rather than considering this to be a drawback, let us instead see it as nothing more than the decision making process of a collective, based on previous experience and a healthy skepticism about new and sometimes untested methods and technologies. Like a double-edged sword, the same cultural bias might catalyze the adoption of the technology as well. This can be seen in the early problems faced during adoption of disruptor technologies like the radio and aircraft. In order to correctly identify stakeholders, it is imperative to accurately weigh the cultural bias associated with a given society; especially since the effects of the technology might be experienced only by a certain section of that society.
- Value System:

  Value systems play a big part in helping the acceptance of technology. Say for example, in the United States the Apple vs. FBI debate has tested the value systems of the public, where protection of individual liberties is of paramount importance. If, however, a similar case would have taken place in the People's Republic of China, the prevailing value systems would have ensured massive public support for the government's stand. In short, the classification of stakeholders may change dramatically. In one society, the Government is seen as a representative of the public, where as in another, the individuals making up that public take on an identity of their own. In one society, people expect a very particular type of privacy from the society they live in, and when that privacy is violated, they experience harm. In a society where the social contract is different, where people have different expectations of privacy, they may not perceive such a harm if they find themselves in the same circumstances.
- Economic Status:

  The economic status of a society broadly refers to the distribution of wealth in the society. We see vastly contrasting structures in Developed Countries and in Developing Countries. One example that could elucidate this difference is the adoption of automation of jobs such as customer service. In developed countries, these are considered tertiary level jobs and do not form the backbone of the economy. Hence the effect would be largely mitigated, and with government intervention, could be avoided nearly entirely. But in the case of developing countries, where semi-skilled jobs like freight services employ a large number of people, we would see a larger effect in terms of unemployment. Accounting for this would play a large part in correctly weighing each stakeholder.

B. General Classes of Stakeholders

Taking into account the effect of automation over the entire cross-section of society, we think it is prudent to identify certain classes of stakeholders which would be present across any situation, regardless of the geographical location, economic status and value system. We do not posit that these general classes stand alone, but rather that an examination of the interplay between the interests of these classes is an important first step for any stakeholder analysis we wish to undertake. It is important to note that these divisions are our own taxonomy, embraced because it is useful for our purposes, not because it is necessarily representative at all levels. We emphasize that the due diligence

required to identify stakeholders must be done in order to ensure the holistic view we would like to achieve in this part of the process.

- Public:
  The public refers to society at large who will inevitably be directly or indirectly affected by the implementation of the automation. Typical issues include job/labor concerns, concerns regarding misuse against other members of the public, and concerns regarding laws enforcing the ethical use of the same, just to name a few. By far this is one of the larger classes, and within it many more might emerge, especially when the technology is targeted to a certain section/strata of society. It is therefore essential to identify which section might have a larger stake in the implementation of the automation, be it positive or as it is often seen, negative.
- Government:
  Government refers to the regulatory body that oversees the ethical use of the automation. We can say in certain scenarios, the lines discerning public interest and government interest might blur, we considered these to be exceptional cases where we can identify one class to represent the interests and stakes of another(as mentioned in the example above). It can be seen in most modern societies today, the public does not necessarily gain adequate representation of their interests through government and similarly, the Government at times ensures the security and well-being of the state even if that does not represent what the public deems fit. For this and reasons that outline similar cases of dichotomy, we would like to represent this as a separate class.
- Corporate Interests:
  Automation today, as it has ever been in the past, is driven by commercial interests. We represent these interests through the class of corporate stakeholders. These would represent the individuals as well as organizations which have driven the case for the automation and would benefit in some monetary fashion by the said implementation. Apart from this, we would like classify industries in this class of stakeholders as well, because of the widespread transmission of ideas and technology, practices linked with an increase in productivity would be widely adopted even if that is at the cost of jobs. Although it would be easy to classify corporate interests as strictly monetary, but in certain scenarios we see that this stakeholder has aligned interests with another classes(such as Government) and hence would require a somewhat different approach to weighing their stakes.

C. Defining a Coherent Value System

Subsequent to identifying stakeholders and classifying the stakeholders into the basic classes, or new ones depending on the circumstances, we must elucidate upon a value system that would act as a guide and in the process allow us to reach a conclusion regarding the ethical implementation of the automation in question.

We propose a negative utilitarian framework focused on preventing harm, where we define harm in terms of violation of a number of values that we consider to be important. We apply this framework with the caveat that its usefulness may break down at certain moral extrema, which result in a reluctance to make trade-offs between particular harms in particular consequences. We identify a rough taxonomy of harms as violations of the following values for stakeholders in an ethical analysis:

- Autonomy of Self
  Autonomy of self, refers to the basic human rights as laid down by the Universal Declaration Human Rights [25], which includes but is not limited to, Right to life, Right to freedom of expression
  As laid-down in the Charter, any harm to these basic rights can considered a violation of our framework, and the person responsible for the analysis should eschew any automation that endangers these basic rights, or in the worst case to mitigate these violations through implementation of other practices that would ensure protection of these rights to the highest degree possible.

A vital point within this tenet is the Right to Privacy, although most people believe this is a fundamental right, in a majority of cases this would take precedence over any technology that infringes on this right. As we are seeing the battle rage between Apple and the FBI, the protection of this entire section of rights would come under the social contract between the individual/s and society at large. There is requirement to state the importance of the protection of these rights, because sometimes these rights may not be protected by the social contract mentioned earlier.

- Welfare/Well-Being

  The ability to ensure Welfare and well being of self, family and society at large has long been considered the highest responsibility. The world we live in today, this is directly associated with sustenance as well as health.

  Any violation of the welfare of the stakeholders through either contamination of sources of sustenance or loss of livelihood could be considered a violation of this principle. Understandably, there are costs to progress, although this framework aims to minimize the infringements, we propose that on finding a violation, action should be taken against the party to ensure that the cost of implementing the automation would be higher than the gain from such technologies, which are implemented by violating these rules.

  The purpose of this framework is less to take a moral stand but rather a way to balance the effect of consequences on the stakeholders, by suggesting action we believe that the framework used in conjunction with the ground reality would help the user come to an understanding of how to ensure a balance is reached.

- Ecocomic Well-Being

  The above points form moral values we wish to preserve in the course of any decisions made within the context of our moral framework. In the course of so doing, we believe it is important to note that many harms or benefits to the welfare and autonomy of individuals we must consider are economic in nature. That is, they may be considered in terms of costs leveraged upon their persons, taxes leveled against corporations, tax incentives provided to farmers that grow particular crops, and other similar structural forces. Therefore, when attempting to analyze harm, it would behoove us to consider the economic outcome, and how it may inform the other values we wish to preserve in our efforts to make ethical decisions. Economics, then, can be viewed as a sort of ethical algebra: a common language that we may use to describe and compare different ethical scenaios in order to discover the the solution that produces economic well-being. In this sense, we might consider economic well-being to be the state in which people are able to fulfill their needs as stated above, unhindered by forces of debt or poverty.

VI. Self-Driving Cars: an Application of our Framework

A. Background

Long a fascination for inventors, the seed for autonomous cars was sewn in the 1960s with the Stanford Cart [26]. The cart navigated small spaces by taking photos of its surroundings, using a simple computer program to analyze the photos, and choosing a path to travel depending on that analysis. Two decades later, a vision-guided Mercedes-Benz van designed by Ernst Dickmanns navigated empty streets at speeds of up to 39mph. By 2009, Google began privately testing anonymous vehicles on closed streets, and by 2012, changes in various states' laws allowed for testing on city streets. As of 2016, Google claims over 1.3 million autonomous miles driven, although Google's cars are limited to 25mph and must always have a human driver present. [27] As driverless car technology improves and becomes more widely implemented, the true societal impact of this technology will begin to be felt.

B. Application

We're in an interesting time regarding autonomous vehicle technology. It's clear that the technology will become widely used, but it has yet

to actually happen. This presents an opportunity to consider how, as a society, we should approach the ethical ramifications of this potentially society-altering technology. To do so, we'll apply the ethical framework laid forth earlier in this paper to autonomous cars.

To begin, it is important to identify the three main stakeholders in driverless car technology: the public, the government, and corporations. These stakeholders are universal, but the beliefs of each stakeholder can vary depending on cultural biases, value systems, and economic status. Also, the weight given to each stakeholder can vary based on these factors. In the interest of concision, this application will focus on a generalized idea of American cultural bias, values, and economic status.

Because the United States is such a demographically, economically, and culturally diverse country, it is difficult to make generalizations of the country as a whole. That being said, it is possible to extrapolate mainstream value systems of our three stakeholders (public, government, and corporations). The American public values individualism and personal rights. Corporations value free market principles and the acquisition of capital. The government acts as a mediator (sometimes poorly) between the interests of corporations and of individuals. Each stakeholder has the power to act on the other and bring about change. With the stakeholders, value systems, and power structure identified, it is possible to apply our ethical framework to autonomous vehicle technology in the United States.

Autonomous vehicles have a lot to offer a utilitarian. A recent Virginia Tech study [28] shows that the crash rate of Google's self-driving cars is lower than the national crash rate of conventional cars. Google's cars experience 3.2 crashes per million miles while conventional cars experience 4.2 crashes per million miles. With autonomous vehicle technology continually improving, the technology will become even safer in the years to come. This sounds great, but is a full on, no-strings-attached embrace of the technology the best course of action? For a utilitarian, the answer may be "yes", but there is a problem with this approach. Autonomous vehicles will face a variety of moral dilemmas, and a one-size-fits all ethical framework doesn't cover all of these possibilities sufficiently.

Consider the following example: during the November 2015 terror attacks in Paris, ride-sharing app Uber's algorithmic design dramatically raised prices in the areas near the attacks due to an increase in ride requests from people trying to flee the chaos. In this situation, stakeholders (public and corporate) were in direct opposition to each other. Uber saw the situation as a moneymaking opportunity, while the public saw Uber as price gouging and profiting from a disaster. Eventually, Uber gave in to public demand and lowered prices. Situations such as this will become the norm in the age of autonomous cars. In fact, Uber is aggressively seeking driverless car technology for use in their taxi company.

In the case of Uber raising the prices on people fleeing a disaster, utilitarianism supports profit maximization as a means of supporting the greatest societal good. Clearly, this is a morally unfavorable position for the public, but fortunately, the moral framework set forth in this paper allows an exception to utilitarianism on the basis of protecting an individual's welfare and well-being. In the case of the Paris attacks or hypothetical future situations involving Uber autonomous vehicles, our moral framework prioritizes lives in danger over profits in dangers.

Consider another hypothetical example: you're riding down the highway in your autonomous car and a large object falls off of a truck in front of you. Your car cannot stop itself in time and must decide on the best course of action. If it continues straight and runs into the object, it places your life in serious jeopardy. There are two motorcyclists on either side of your car; one is wearing a helmet and the other is not. Swerving into one of the motorcyclists will likely save your life, but jeopardize the life of one of the motorcyclists.

With strict utilitarianism, your car would likely swerve into the motorcyclist wearing the helmet because it creates the best chance of survival for all involved, but is it fair to punish a motorcyclist for being responsible and wearing a helmet? What

are the legal and moral ramifications of such a programmed decision? In this scenario, our ethical framework would fall back onto the autonomy of the individual. They could decide how their car is programmed to react in cases such as these. Alternatively, self-driving cars could rely on an element on randomness in making such decisions or even cede control to a human driver (local or remote) despite the car's ability to handle situations such as this without human intervention. The important point is leaving autonomy in the hands of the individual, not the machine.

While far from foolproof, our ethical framework allows for important exceptions to utilitarianism. The implementation of driverless cars will undoubtedly lead to many unique moral dilemmas, therefore the existence of *fallback rules* to a society's moral framework will be critical. In the previous examples, the *fallback rules* of our ethical framework placed health and well-being above financial profit and the autonomy of a human over the autonomy of a machine. These are only two examples of the multitudes of moral dilemmas that will become ever-present in a driverless car future, and it is critical that we continue to work toward developing an ethical framework to deal with every possibility.

VII. Conclusion

Rise of the automated tools designed to accomplish certain tasks and ethical problems stemming thereof is a very potent area for dialogue. It is evident that the set of current solutions have more cons than pros when one put them under scrutiny. We claim that the framework that we proposed in this paper would minimize the short comings and maximize the gains. The efficacy of the proposed solution was evident in the given use case about self-driving cars. We have showed that the proposed solutions would safeguard the autonomy of self, welfare/well-being, and the economic goals of the relevant stakeholders in such a way the maximum utility is obtained while minimizing the harm. It is worthy to note that, while competent, at a certain point when a certain threshold is exceeded we are still yielding the control to a human. The very existence of this option may help some reluctant individuals to reduce their resistance towards automation because, humans no matter how illogical it is, tend to think that if they have control they can fix any problem despite the fact that an autonomous system deemed a feasible solution does not exist. But one must always remember that until we do obtain a certain level of automation, real the social impact of such a technology can only be speculated. Minute details that escape the most prudent analysis might have very far reaching consequences. And thus we stand on the precipice and stare into the void, and does the void stare back into us?

References

[1] Homer, *Iliad*, 1194 BC, Chapter 18. 371 ff.

[2] Aaron J. Atsma, "Automatones," 2015, Available: http://www.theoi.com/Ther/Automotones.html , Accessed 2016-03-13.

[3] Lie Yukou, *Liezi*, pp. 83–86, 600 BC.

[4] Sarah L. Higley, "Alien intellect and the roboticization of the scientist," *Camera Obscura*, vol. 14, no. 1-2 40-41, pp. 129–160, 7 1997.

[5] "Google self-driving car project," 2016, Available: https://www.google.com/selfdrivingcar/ , Accessed 2016-03-13.

[6] "Your autopilot has arrived," 2016, Available: https://www.teslamotors.com/blog/your-autopilot-has-arrived, Accessed 2016-03-13.

[7] I. Liu and B. Ramakrishnan, "Bach in 2014: Music Composition with Recurrent Neural Network," *ArXiv e-prints*, Dec. 2014.

[8] James Gould-Bourn, "Man combines random people's photos using neural networks and the results are amazing," 2016, Available: http://goo.gl/11Lg2W, Accessed 2016-03-13.

[9] Justin Mahboubian-Jones, "Fully-automated security robots have arrived," 2016, Available: https://www.pixeldynamo.com/news/2014/11/18/39782/real-life-robocop-sorts-way/, Accessed 2016-03-13.

[10] Eyal Reingold, "The Turing Test," 1999.

[11] Arthur C. Clarke, *Profiles of the Future*, chapter "Hazards of Prophecy: The Failure of Imagination", Phoenix, 1962.

[12] Michael Sainato, "Stephen hawking, elon musk, and bill gates warn about artificial intelligence," 2015, Available: http://observer.com/2015/08/stephen-hawking-elon-musk-and-bill-gates-warn-about-artificial-intelligence/, Accessed 2016-03-13.

[13] Luke Muehlhauser, "What is agi?," 2013, Available: https://intelligence.org/2013/08/11/what-is-agi/, Accessed 2016-03-13.

[14] Peter Gaskell, *The manufacturing population of England, its moral, social and physical conditions*, Baldwin & Cradock, 1833.

[15] Karl Marx, "The strife between workman and machine," in *Capital*, vol. 1, chapter 15. 1867.

[16] Cade Metz, "GoogleâĂŹs ai takes historic match against go champ with third straight win," 2016.

[17] John Markoff, “Computer wins on âĂŸjeopardy!âĂŹ: Trivial, itâĂŹs not,” 2011.
[18] Aaron Smith, “Public predictions for the future of workforce automation,” 2016.
[19] Samuel Gibbs, “Musk, wozniak and hawking urge ban on warfare ai and autonomous weapons,” 2015.
[20] Ray Kurzweil, *The singularity is near: When humans transcend biology*, Penguin, 2005.
[21] Bill Joy, “Why the future doesnâĂŹt need us,” *Nanoethics. The Ethical and Social Implications of Nanotechnology*, pp. 17–30, 2000.
[22] Jaron Lanier, *Who owns the future?*, Simon and Schuster, 2014.
[23] Jameson M Wetmore, “Amish technology: Reinforcing values and building community,” *Technology and Society Magazine, IEEE*, vol. 26, no. 2, pp. 10–21, 2007.
[24] Friedrich Engels and Tristram Hunt, *The Origin of the Family, Private Property and the State*, Penguin UK, 2010.
[25] UN General Assembly, “Universal declaration of human rights,” *UN General Assembly*, 1948.
[26] Hans P Moravec, *The Stanford cart and the CMU rover*, Springer, 1990.
[27] Alex Davies, “Google’s self-driving car caused its first crash,” *Wired*, 2016.
[28] Myra Blanco, Jon Atwood, Sheldon Russell, Tammy Trimble, Julie McClafferty, and Miguel Perez, “Automated vehicle crash rate comparison using naturalistic data,” Tech. Rep., Virginia Tech Transportation Institute, 2016.