Department of Political Science

Course of Political Philosophy

The Ideal Utilitarian Agent in The Age of New Technologies: How Utilitarian Logic Challenges Our Exclusivity as Moral Persons

THESIS SUPERVISOR

Prof. Sebastiano Maffettone

CANDIDATE

Emanuele Pesoli

ID: 094902

Academic year: 2022/2023

Table of Contents

Introduction *3*

1. Utilitarianism: A Flawed Moral Theory *5*

1.1 The Origins of Utilitarianism **5**

1.2 The Three "Great" Utilitarians: Bentham, Mill, Sidgwick **7**

1.3 What Is Utility? **10**

1.4 The Challenges of Consequentialism and Welfarism **13**

1.5 Is Utilitarianism Still Relevant Today? **17**

2. Technology: Friend or Foe? *19*

2.1 Social Media: A Cautionary Tale **19**

2.2 Artificial Intelligence: A Hazier Picture **21**

2.3 The Next Step: Brain Computer Interfaces **25**

2.4 The Posthuman Question **27**

2.5 The Utilitarian Position **31**

3. Greene's Trolley Problems *34*

3.1 The Three Scenarios: Switch, Footbridge, Loop **34**

3.2 A Conditional Aversion to Violence: The Human Reaction **35**

3.3 Torn Between Utilitarianism and Asimov's Laws: The Robots' Reaction **38**

3.4 A Speculative Mystery: The Posthuman Reaction **40**

Conclusion *43*

Bibliography *47*

Introduction

The world we live in is constantly changing before our very eyes. Technology has made things our ancestors could have never dreamt of, possible. But will technology help us overcome future challenges, or will we become more and more dependent on it? Would we become slaves to our own creations? Perhaps, we should all have a moral compass, something that would help us navigate the turbid waters of this new, frightening, technological world without losing track of our principles, our dignity as persons, and, most of all, our humanity. This paper presents utilitarianism as a potential candidate for said compass. But, why utilitarianism?

Certainly, it would not be for the strength of its arguments or their consistency, as we will see in the first chapter of this paper. The main reason why I chose utilitarianism is because, despite its arguments being conflicting and overly complex, it still is attractive to us, much to some theorists' frustration. Judith Thomson, for instance, claims that utilitarianism "keeps on reappearing every spring, like a weed with long roots[1]"; its main attraction, at its core, lies mainly in that it "appeals to our benevolence" and that it "[…] values welfare and nothing else[2]".

Of course, the goal of maximizing human welfare at all costs has some serious and problematic ramifications, but I would argue that analyzing these ramifications, and the moral dilemmas that come with it, may very well help us understand how far we are willing to push ourselves for the sake of the greater good, and how much trust we should put in the hyper-intelligent machines which will one day become an important part of our daily routines. This paper is not a critical analysis of utilitarian literature. It merely gives a brief glance to this moral theory's main arguments because they reflect something deeper, something uniquely human, they are 'rooted', echoing Thomson, in our intuitions, "[…] a form of knowledge that appears in consciousness without obvious deliberation[3]". However, because of the immediacy of intuitions, they often conflict with any reasonable judgement. The very reason why utilitarianism is appealing to us, our intuition that human welfare matters, seems, then, to directly cause its failure as a moral theory. This brings us to the second chapter of this paper.

[1] Judith Jarvis Thomson, "Goodness and Utilitarianism", *Proceedings and Addresses of The American Philosophical Association*, 67, No 4(Jan., 1994) **as cited in** Robert Shaver, "The Appeal of Utilitarianism", *Utilitas,* 16, No 3(Nov., 2004), 235.
[2] Robert Shaver, "The Appeal of Utilitarianism", *Utilitas*, 16 No 3(Nov., 2004), 236.
[3] "Intuitions", PhyscologyToday, accessed July 18th, 2023 https://www.psychologytoday.com/us/basics/intuition

The main utilitarian texts that will be presented in this paper date back to the 17th, 18th, and 19th century; most of the authors and theorists that will be presented and discussed hereafter could not have ever imagined what the future would bring, and how certain technological innovations might have posed a significant challenge to their arguments. Many of the things they would regard as unquestionable truths may very much appear to us as easily objectionable. For instance, not one of the many utilitarian authors that will be presented in this paper (some of which are still living today) has ever questioned the fact that a utilitarian agent could not be human. Would utilitarianism fail as a moral theory if the agent in question was, for instance, a robot? Could said robot become the perfect utilitarian agent, maximizing the collective utility of the entire world, unburdened by the weight of human intuitions? Artificial intelligence (AI) is becoming more and more proficient in solving complex problems faster and more efficiently than any human can. Should we trust these robots with our lives? Should we treat them as our equals on a moral level? Perhaps. As mentioned above, however, we should not dismiss our intuitions: we cannot allow ourselves to lose track of this innately human quality. So, maybe, the answer is not to look at technology, but rather to find a way to improve ourselves biologically, so that we may be able to maximize utility more efficiently.

This paper's main goal is to understand, using utilitarian literature, and the dilemmas and contradictions that come from it, whether maximizing humanity's welfare as a whole (and what it would take to do that) is compatible with the imperfect reality of being human and, if not, how distant from us would the 'ideal moral agent' have to be to perform said utility maximization; should we allow our faith to be in our hands, in the hands of machines, or in the hands of an enhanced version of ourselves? How does utilitarianism challenge our preconceived notions of what makes someone or something a moral person?

1. Utilitarianism: A Flawed Moral Theory

1.1 The Origins of Utilitarianism

Although utilitarianism as a moral theory developed in the 1700s and 1800s in Europe, the origins of its core principles can be traced back to ancient China. Mozi (490-403 BC), a Chinese philosopher, was the first to argue that certain customs or traditions shouldn't simply be accepted as given but, rather, should be assessed against a certain standard, to understand if they yield more harm than good; if that's the case, then they should be abandoned. Other famous thinkers of the time, such as the Indian Gautama, universally known as Buddha, and later the Greek Epicurus, shared similar beliefs; in particular, Gautama stressed how important it was to limit the suffering of others as much as possible, while Epicurus' conviction that the standard of rightness was inextricably linked with the pleasure and pain axis would prove to be one of the core pillars of early utilitarians[4].

Modern Europe's interest in the concept of adopting as a standard of rightness the general good, or happiness of the people, picked up steam in the 17th and 18th centuries, with works such as Cumberland's *De Legibus Naturae* (1672) giving rise to a primitive form of utilitarianism; theological utilitarianism[5]. Cumberland (1631-1718) firmly believed that, if God was benevolent, then he would have naturally wanted all the creatures that inhabited the world he created to be as happy as they could; the collective good of all these individuals that live in the world is described in the *De Legibus Naturae* as a whole whose value is greater than the sum of its parts. To maximize this collective good, Cumberland argues that men should simply abandon themselves completely to divine providence, which, translated in earthly matters, means that they should respect the laws and systems in place, as they are themselves an expression of God's will. Later authors, such as Hutcheson (1694-1746), distanced themselves from theological arguments and argued that any rational human agent possessed a moral sense; this pushed them to act in such a way that maximized the happiness of the community[6]. In 1726, Hutcheson published *An Inquiry into the Original of our Ideas of Beauty and Virtue* in which the phrase '*the greatest happiness of the greatest number'* first appeared. These words would later be quoted by different authors in the following years including the Italian Cesare Beccaria, from whom

[4] Katarzyna De Lazari-Radek and Peter Singer, *Utilitarianism: A Very Short Introduction* (Oxford: Oxford University Press, 2017), 1-2.
[5] Piergiorgio Donatelli, *Le storie dell'etica: Tradizioni e problemi* (Roma: Carrocci Editore, 2022), 63-64.
[6] Ibid., 64-68.

Jeremy Bentham (1748-1832), widely regarded as the founder of utilitarianism, read, and adopted it from, as a sort of motto to summarize his main arguments[7].

[7] Radek and Singer, *Utilitarianism*, 2-3.

1.2 The Three "Great" Utilitarians: Bentham, Mill, Sidgwick

Among the many utilitarians who lived throughout history, three men are particularly noteworthy: Jeremy Bentham (1748-1832), John Stuart Mill (1806-1873) and Henry Sidgwick (1838-1900).

Bentham, an Oxford alumnus, became known across Europe for his reform proposals of legal systems and prisons; for instance, his '*Panopticon*', a prison where prisoners could work while being constantly observed, made waves across the Old Continent. In 1780 he published the *Introduction to The Principles of Morals and Legislation*, considered by many as one of the most important utilitarian texts to date, instrumental in his pursuit of creating an ideal code of law, something he dedicated himself to for the last two decades of his life[8]. Bentham was thus way more ambitious than his predecessors, as he sought to apply the utility principle, which he also referred to as the greatest happiness principle, to various fields, including politics and the law; he argued for a more democratic and egalitarian version of utilitarianism, which would maximize the utility of a community by managing a fairer distribution of resources between its members[9].

John Stuart Mill was the son of one of Bentham's followers and friends, James Mill, and a child prodigy who by the age of eight could read both ancient Latin and Greek and by fifteen had competences in every major academic discipline; it was around this time that he gained an interest in Bentham's work. Nevertheless, his most important contributions to utilitarian literature, *On Liberty* (1859) and *Utilitarianism* (1861) were published much later in his life[10]. Mill, though certainly inspired by Bentham, developed his own version of utilitarianism, which was more focused on everyday morality and was slightly more sensible to individual needs. In *On Liberty*, for instance, he argues that public authorities' interference in citizen's lives should be limited as much as possible. Furthermore, he also went as far as to claim, in *Utilitarianism*, that justice should merely ensure the essential conditions of humanity's well-being, without threatening its freedoms[11]. It's only natural, then, that Mill's justification of the utility principle would feature a stronger focus on the individual good.

[8] Radek and Singer, *Utilitarianism*, 3-5.
[9] Donatelli, *Le Storie dell'etica,* 71-74.
[10] Radek and Singer, *Utilitarianism*, 7-9.
[11] Donatelli, *Le Storie dell'etica*, 74-75.

He argues:

> No reason can be given why the general happiness is desirable, except that each person, so far as he believes it to be attainable, desires his own happiness.[…] each person's happiness is a good to that person, and the general happiness, therefore, a good to the aggregate of all persons[12].

Here we see, basically, an opposite version of Cumberland's theory, which was detailed in section 1.1. Instead of starting from a self-evident truth (in Cumberland's case, that God is benevolent) and from that deducing that the ultimate goal is the maximization of everyone's happiness, Mill adopts an inductive approach; since we know from experience that an individual desires their own happiness, then everyone else must do the same. Thus, promoting the happiness of the community (intended as an aggregate sum of individual utilities) will yield good to all. The main issue that seems to escape Mill is that he gives for granted that all these individual utilities aren't mutually exclusive (i.e., an egoistic individual's happiness may affect other people's happiness in a community) [13].

A notable critic of Mill was Sidgwick, an academic who dedicated his whole life to improve and update his first and most important work, *The Methods of Ethics* (1874); in it, he presented three different methods of reasoning (egoism, utilitarianism and intuitionism) with particular care, so much so that many believe Sidgwick to be one of the first, if not the first, to adopt a comparative method of study for a philosophical work, something that has become the norm now[14]. While Mill was a proud member of the inductive school of thought, meaning he believed that we learn what is morally right or wrong from our own life experiences, Sidgwick belonged to the intuitive school, and firmly stood by the conviction that ethical self-evident principles indeed exist, and that we can't possibly learn what is right or wrong from experience nor from everyday morality. Even supposedly universal rules such as '*don't tell lies*' may sometimes conflict with utility maximization, and thus exceptions would have to be made (i.e., white lies) depriving them of their self-evidence. After years of research, Sidgwick identified three self-evident principles: *justice* (always treat similar situations as if they were the same), *prudence* (give every moment of an individual's existence the same weight) and *benevolence* (treat an individual's good like any other individual's good). Still, Sidgwick was not

[12] Radek and Singer, *Utilitarianism*, 21.
[13] Ibid., 22.
[14] Radek and Singer, *Utilitarianism,* 11-13.

unaware of a potential flaw in his reasoning; mainly, that egoism could prove to be a very attractive alternative to the self-evident principle of benevolence, thus creating a major inconsistency within his theory[15].

[15] Radek and Singer, *Utilitarianism,* 23-27.

1.3 What Is Utility?

Bentham, Mill, Sidgwick, Cumberland, and Hutcheson all had their differences, but they at least all agreed on what utility is: they all believed that the "experience or sensation of pleasure is the chief human good"[16], making them hedonists[17]. Bentham, in his *Introduction to The Principles of Morals and Legislation,* provided a very simple explanation of the hedonist doctrine, which all the other authors discussed so far could also easily get behind:

> Nature has placed mankind under the governance of two sovereign masters, *pain*, and *pleasure*. It is for them alone to point out what we ought to do, as well as to determine what we shall do[18].

The importance of pain and pleasure in regards to human agency was recognized since ancient times(as mentioned in section 1.1), but both Bentham and Mill endorsed a more complex version of hedonism, which saw pleasure becoming more than a simple physical sensation; they both saw pleasure as a "[…] mental state or property that is or that has a certain something that is 'what it is like' for its subject; a certain feel, feeling, felt character, tone or phenomenology[19]".
This millennial tradition, however, was challenged by Robert Nozick and his 'experience machine'; he argued that scientists could connect an individual to a machine which constantly pumps drugs into their system, making them feel every positive emotion or mental state they could ever experience in their lives (falling in love, being rewarded for an accomplishment etc.). Still, that person would have to give up the chance of actually experiencing real emotions and pleasure to live attached to a machine, something that hardly anyone would choose[20]. This is the first example where technology becomes a hindrance to utilitarian reasoning, something that will be discussed at length in subsequent chapters. For now, what is important to note is that the hedonistic definition of utility is unsatisfactory as it contradicts Bentham's *'greatest happiness'* motto; if the sensation of pleasure (either mental or physical) is what it takes to maximize utility, then, even without considering Nozick's machine, everyone would start injecting themselves with drugs and become addicts, and no one would be happy.

---

[16] Will Kymlicka, *Contemporary Political Philosophy: An Introduction* (Oxford: Oxford University Press,2002),13.
[17] Radek and Singer, *Utilitarianism*, 42.
[18] "Hedonism", The Stanford Encyclopedia of Philosophy, last updated October 17th 2013 https://plato.stanford.edu/entries/hedonism/#EthHed
[19] Ibid.
[20] Kymlicka, *Contemporary Political Philosophy*, 13-14.

Another strand of utilitarian thought provides a viable alternative to hedonism and is strongly linked to preference satisfaction: basically, since people's preferences are an expression of their desires, which are indicative of what will yield good to them, satisfying their desires would maximize their individual utility, and because of that, the more individual preferences are satisfied, the more collective utility will be maximized[21]. Although preference utilitarians avoid Nozick's trap by leaving the feeling of pleasure out of the equation[22], many objections can still be raised against them. For instance, consider a slave who wants freedom but, as he grows older, gradually becomes accustomed to his condition, and doesn't want to be freed anymore; or a student who romanticizes being a lawyer, but once he starts studying law realizes that he does not have the qualities required for said profession[23]. The solution to the problem posed by these preferences, which can be referred to collectively as irrational, according to some preference utilitarians, is to simply exclude them from utility calculations and just focus on those preferences that we express only when we are "fully informed and thinking clearly"[24]; which are very difficult to define and/or measure and may lead to further problematic implications. For instance, should someone who desires nothing more than to follow God's word not go to mass because they would miss a nice sunny morning in the countryside? The answer that preference utilitarians would give here is that they would choose not to go to mass if they were fully informed (if they understood that there's no proof regarding God's existence, and no way to tell if praying to him would yield any benefit to them or to anyone else) but still, they wouldn't be happier because their real, present preferences have been disregarded[25]. Preference utilitarians 'argument according to which satisfying informed preferences will yield the most good is more convincing than what hedonism offered, at least on paper, but once one starts diving beneath the surface a lot of problems, mostly related to the semi-impossibility of discerning an irrational preference from a rational one, in which they are expressed, arise[26].

Beyond hedonists and preference utilitarians, there is yet a third group that has another interpretation of what utility is and how we can maximize it: they are known as ideal utilitarians, because they believe that some ideals (i.e., knowledge, beauty, freedom) do possess an intrinsic value independently of our preferences and can thus make our lives better the more of them we have, regardless if we desire them or not[27]. Ideal utilitarians technically fix the problem of the *'contented*

[21] Kymlicka, *Contemporary Political Philosophy*, 14-15.
[22] Radek and Singer, *Utilitarianism,* 47.
[23] Kymlicka, *Contemporary Political Philosophy,* 15-16.
[24] Radek and Singer, *Utilitarianism,* 50.
[25] Ibid, 50-51.
[26] Kymlicka, *Contemporary Political Philosophy,* 18.
[27] Radek and Singer, *Utilitarianism,* 52-53

slave' mentioned earlier; even if the slave consciously loses their desire to be free, this doesn't take away from the fact that if they gain freedom their life will automatically improve, because of the intrinsic value of freedom as an ideal. Indeed, even Mill, in his work *On Liberty* (1859), defended freedom and regarded it as a prerequisite for any person's happiness; still, ideal utilitarians do not have a set of ideals set in stone, like Sidgwick did, and if one tries to test their self-evidence, the results will not be satisfactory. For instance, if we take freedom as an ideal of intrinsic value, this means that we should allow people to be free above all else, even if they might hurt themselves or others in pursuing it (i.e., allowing people not to wear seatbelts, or drive intoxicated). Similarly, if we take truth as an ideal, it would mean that we can't ever lie in our lives; something that will inevitably hurt others and affect our personal relationships[28].Sidgwick seems to have formulated a theory about self-evident principles better than ideal utilitarians. *Justice*, *prudence*, and *benevolence*, as formulated by him at least, remain very difficult to argue against as universal values (section 1.2).

Trying to answer the question that titles this section isn't easier now than it was more than two centuries ago; even with two-hundred years of literature at our fingertips, the best we can do is to look at the lowest common denominator between all the theories considered thus far and state that, for utilitarians, maximizing utility means maximizing a "person's conception of their own well-being"[29].

[28] Ibid, 53-56.
[29] Amartya Sen, "Utilitarianism and Welfarism", *The Journal Of Philosophy*, 76 no 9(Sep., 1979): 463.

1.4 The Challenges of Consequentialism and Welfarism

In the last section of this paper some of the contradictions that stem from the main strands of utilitarian thought were presented; for this very reason it is perhaps relevant to analyze consequentialism, to which utilitarianism is inextricably tied to[30], before diving deeper in the contradictions and problematic implications of utilitarianism itself. Consequentialism predates utilitarianism and is arguably more intuitive: what consequentialists say is that the better action is that which will yield the better consequences[31]. As a matter of fact, some argue that utilitarianism is "consequentialism plus hedonism"[32], in the sense that it adds happiness or pleasure as a variable when calculating the consequences of an action. Indeed, consequentialism was first endorsed officially by Cumberland as the only true criteria to judge moral actions[33]. It would be a mistake to assume, however, that consequentialism is somehow less flawed than utilitarianism; on the contrary, it could be argued that some of the biggest problems of utilitarianism are inherited from consequentialism.

To show exactly why consequentialist logic can be problematic, a very famous example will be considered, that of *Baby Hitler* as first conceived by J. C. C. Smart:

> It is […] 1895 and you are walking across the bridge that spans the river in Linz, Germany. You notice some of the children […] playing by the water's edge. […] one of the smaller boys slips into the water and […] it is clear that he cannot swim and will drown if you do not save him. You jump in the rushing water and […] swim him to safety. […] you saved the six-year-old son of Klara Hitler, Adolph[34].

Of course, the action of saving a drowning child seems to have, intuitively, the better consequences of letting him drown. Indeed, if the action was evaluated twenty years later, it would still be considered a good action; however, if what you did was evaluated in 1945, it could be seen as a crime

[30] Sen, "Utilitarianism and Welfarism", 463.
[31] Jared Rudolph, "Consequences and Limits: A Critique of Consequentialism", *Malcaster Journal of Philosophy,* 17 no 1(Jan., 2011): 65-66.
[32] Ibid.
[33] Donatelli, *Le storie dell'etica*, 66.
[34] J.C.C Smart, "Extreme and Restricted Utilitarianism", *The Philosophical Quarterly*, 6(1956) **as cited in** Jared Rudolph, "Consequences and Limits: A Critique of Consequentialism", *Malcaster Journal of Philosophy,* 17 no 1(Jan., 2011), 68-69.

worthy of capital punishment[35]. Consequentialists need to clarify the time of the action's evaluation and the scope of the act itself, that is, to what extent the person is responsible for all the possible future and (unpredictable) consequences of a seemingly good action in the present, if they want their theory to be seriously considered as an ethical theory[36]. Thus, consequentialism's main issue lies in just how much agency, or lack thereof, the person who acts has or, more specifically, of "what is or isn't in a person's control"[37].Hutcheson seems to take a clear stance on this matter; he argues that all the possible consequences of an action, even those that seem not to have any effect on the present, should be taken into consideration[38]. On the other hand, Mill seems to understand the limited scope of human agents, and thus argues that actions are right "in proportion as they tend[39]" to yield happiness.

The same issues stemming from consequentialist logic are very much present in a variation of utilitarianism, known as welfarism, which claims that:

> The judgment of the relative goodness of alternative states of affairs must be based exclusively on, and taken as an increasing function of, the respective collections of individual utilities in these states[40].

On the surface this seems reasonable, and perfectly in line with utilitarian logic; but, perhaps, this is welfarism's greatest flaw.

For example, consider a poor policeman(*p*) and a rich dreamer (*r*) and two scenarios: *x* where there are no taxes, and *y*, where *r* is taxed in favor of *p*. In this case utilitarians, as well as the average person, would likely choose the *y* scenario; the overall utility distribution is fairer, and the conditions of the disadvantaged party improve. Now, welfarism concedes that we can rank *y* above *x* if and only if we always rank any scenario that shares the utility distribution of *y* above one that shares *x*'s. To understand why this might have troubling implications, let us consider another couple of scenarios: *a* and *b*. In *a*, like in *x*, the poor policeman is miserable while the rich dreamer is living his best life; in *b*, however, the policeman tortures the dreamer and gains pleasure from it. Of course, intuitively,

[35] Rudolph, "Consequences And Limits: A Critique of Consequentialism", 69-71.
[36] Ibid.
[37] Sen, "Utilitarianism and Welfarism", 467
[38] Donatelli, *Le Storie dell'etica*, 68.
[39] Ibid., 76.
[40] Sen, "Utilitarianism and Welfarism", 468.

most people would not allow anyone to be tortured, and would thus rank *a* over *b*; however, welfarism demands that if we allow the policeman to benefit from taxes in *xy*, than we must allow him to torture the dreamer in *ab*, because, on utilitarian grounds at least, *y* and *b* are equally valid[41].The dangers of welfarism are, indeed, mostly related to the fact that it asks us to look at the bigger picture and demands that we leave out the empathy we may feel for the tortured dreamer in the *ab* scenarios, just like we let some of his money go to the policeman in the *xy* scenarios.

A milder, more sugarcoated version of welfarism is weak Paretianism, which argues:

> If state of affairs x is higher than state of affairs y in everyone's utility ranking, then x is a better state than y[42].

To explain this quote, another example may be considered.
Imagine there are two readers, one more prudish and conservative (*P*) and one racier and more uninhibited (*L*) and consider three scenarios: in the *p* scenario the prudish reader reads an explicitly sexual book, in the *l* scenario the racier reader reads the same book, while in the *o* scenario no one reads the book. In this case, the *p* scenario is ranked above the *l* scenario by both the prude and the racy reader; *L* takes pleasure in imagining the horror on the face of *P* while reading the book while *P* could not bear the idea of *L* enjoying the book, thus preferring to read the book himself[43]. While welfarism completely rules out any non-utilitarian variable, weak Paretianism goes out of its way to do the opposite. If we consider again the *xy* and *ab* scenarios it's very much possible that both the rich dreamer and the poor policeman would rank the *y* scenario over *x*, because the dreamer wouldn't mind giving some of his money to the policeman; this, consequently, would make *y* better than *x* on Paretian grounds. However, it's unthinkable that the dreamer would ever accept to be tortured to increase the policeman's utility; because of this, weak Paretianism would fail to conclude what be the best scenario between *a* and *b* would be, as the policeman and the dreamer would rank these differently.

On balance, welfarism does succeed logically, but lacks any attraction on a moral level, because it goes against our intuitions of what is right and wrong; weak Paretianism, on the other hand, fails in

[41] Ibid., 473-474
[42] Sen, "Utilitarianism and Welfarism", 479.
[43] Ibid., 480-481.

the logical department, but appears attractive and reasonable to us because it conforms to our moral intuitions. I believe it relevant, given the context, to briefly mention Richard Hare's distinction between 'level 1 thinking', or *intuitive* thinking, and 'level 2 thinking', or *critical* thinking; he argues that the first level is important during 'fight or flee' situations, where you must take a decision quickly, while the second must be used to solve complex dilemmas that require a more critical approach[44]. Obviously, weak Paretianism is a 'level 1' winner[45] while welfarism seems to be justified only if one looks at the aforementioned scenarios with a more critical lens, that is, if one works on 'level 2' thinking. However, it is also entirely possible for some to carry out an in-depth, critical analysis, and still conclude that welfarism is unacceptable[46] especially if violations of human rights are involved, like in the *ab* scenarios.

[44] Sen, "Utilitarianism and Welfarism", 475.
[45] Ibid., 483.
[46] Ibid., 488.

1.5 Is Utilitarianism Still Relevant Today?

The main appeal of utilitarianism during its early days was the promise of a freer, more balanced society in a time period where the vast majority of the population lacked rights and protection from the very small, privileged elites that ruled over it; at this time, promoting '*the greatest happiness of the greatest number*' seemed like a very progressive and noble objective to achieve, as most people were disadvantaged, disenfranchised or abused by the system they were part of. Nowadays, however, most of the citizens have gained the rights they had asked for and utilitarianism's appeal has lost much of its attractiveness; because the only discriminated groups in modern societies are minorities (i.e., LGBTQI+, indigenous peoples, ethnic minorities), whose interests are often in contrast to those of the wider population, utilitarianism would most likely defend the *status quo* instead of arguing in favor of marginalized groups' rights[47].

Interestingly, in 2012, the United Nations published its first *World Happiness Report*, recognizing happiness as an important criterion to formulate policies[48].Happiness is assessed following either of the following methods: the first seeks to add up all the positive experiences people have had, and then subtract the negative ones from the sum; while the second consists in simply asking the respondent how satisfactory their life has been up to that point. According to the first method, the happiest countries turned out to be African or South American ones, like Nigeria or Mexico, while according to the second richer countries, like Switzerland and Denmark, tended to get on the podium[49]. What clearly emerges from this is that happiness, being subjective, inevitably depends on how one assesses its own and on how one measures it, something that utilitarians learned the hard way.

Utilitarians have tried to solve some of the biggest inconsistencies of their theory by finding new ways to measure utility, which resulted in even more problems and moral dilemmas. In short, because "the winds of utilitarian argumentation blow in too many directions"[50], this moral theory has lost so much of its initial intuitiveness and simplicity that many contemporary utilitarians have argued, paradoxically, that the best way to maximize utility is to abandon the use of utilitarian logic completely[51]. Though it is true that utilitarianism has many glaring and evident flaws, which were extensively discussed in the various sections of this chapter, I would also argue that, in all its different

[47] Kymlicka, *Contemporary Political Philosophy*, 45-47
[48] Radek and Singer, *Utilitarianism*, 115.
[49] Ibid., 116.
[50] George Sher, "Justifying Reverse Discrimination in Employment", *Philosophy and Public Affairs*, 4 no 2(1975**) as cited in** Will Kymlicka, *Contemporary Political Philosophy: An Introduction* (Oxford: Oxford University Press,2002),48.
[51] Kymlicka, *Contemporary Political Philosophy*, 46.

forms, this theory will always be of interest to social researchers, philosophers, and psychologists. Consider, for instance, the questions that can be raised from its literature: are we aware of what actually maximizes our own happiness, or have we '*adapted*' our preferences to fit our social environment? Do we have what it takes to think critically, like Hare preached, and take decisions which go against our natural intuitions, for the sake of collective utility maximization? Do certain ideals have an intrinsic value even if possessing them in excess may lead to more harm than good?

So, to answer the question presented above, yes, utilitarianism can still be relevant today, particularly when it comes to discourses regarding the dangers of new technologies. As mentioned in the introduction, this paper is not interested in analyzing utilitarianism's literature just to prove its validity as a moral theory. The main interest in the utilitarian doctrine is that it reflects the imperfect reality of being human and desperately tries to make it so that everyone is as satisfied with their existence. Of course, the goal utilitarians set for themselves is utopian; you can never really be sure about what makes someone happy or content with themselves unless you'd read their mind. In section 1.3 I argued that the best definition of utility, considering all the interpretations and variations that were given to it over time, was a '*person's conception of their own well-being'*. This definition would be accepted at face value by every author cited above and is, thus, the strongest one to come out of utilitarian literature. But does this definition still hold today? It's hard to say. Many of the authors that were analyzed in this chapter would look at today's world with doubt and suspicion; with big corporations controlling the way people interact with one another in virtual spaces and cameras installed in devices we carry around constantly, some utilitarians may be horrified by the technocratic dystopia we live in.

Is it possible that our desires, our preferences, may be in some way manipulated or coerced into something different without us knowing? Is it possible that the only thing utilitarians seemed to agree on, their ultimate postulate, so to say, could be brought into question? The answer to all this lies right in front of our very eyes; what we need to do now is take a deeper look at the world we live in.

2. Technology: Friend or Foe?

2.1 Social Media: A Cautionary Tale

As mentioned in chapter 1.5, happiness started being used as an important indicator of the quality of life of individuals barely a decade ago, when it was officially endorsed as such by the UN, something that Bentham would be very content with[52]. The main issue with using happiness as a basis for policymaking, or as a unit of measure to understand the average quality of life of individuals, lies in the fact that it is very much subjective and depends on many different factors (the social relationships the person has, how much free time they have or their income). However, according to some researchers, it is entirely possible to measure people's happiness levels by asking them to state how happy they are on a numerical scale from 0(least happy) to 10(most happy)[53], thus making happiness a more statistically viable variable. As technology advanced however, surveys started becoming obsolete in assessing people's happiness. For example, Dodds and Danforth developed a code to evaluate the general level of happiness of a population by carrying out a semantic analysis of millions of tweets (posts on the social network *Twitter*) and associating the use of certain words in them with human emotions, something that was later used in a subsequent study to evaluate the happiness level of various South American countries[54].

Some social media platforms have also been shown to make it easier for people to maintain social relationships, like *WhatsApp*, with which anyone can send thousands of messages a day to their friends and family members, who may live on the other side of the globe. Furthermore, in a survey carried out in Spain, it was observed that people who use social media, regardless of their age, generally appeared to be happier than those who didn't; in particular, elders (65+) who had social media and used it regularly were shown, on average, to have a higher level of satisfaction in their lives than their peers. Overall, it has been shown that people who live more isolated than the average person (not just elders, but also anyone else who suffers from social anxiety and thus struggles to nurture relationships in real-life) greatly benefit from social media, as they can create very complex and fulfilling virtual relationships[55].

[52] Radek and Singer, *Utilitarianism*, 117.
[53] Francisco Mochón, "Happiness and Technology: Special Consideration of Digital Technology And Internet", *International Journal of Interactive Multimedia and Artificial Intelligence*, 5 no 3(Dec., 2018), 163.
[54] Ibid.
[55] Mochón, "Happiness and Technology",165.

Of course, the world of social media isn't as rosy as it would appear; many of these platforms are designed to make users addicted to their content and increasingly more dependent on it.The main problem here lies in the fact that, due to the algorithms behind these apps, people receive personalized content specifically designed to create an immediate and very strong reaction that, more often than not, has the direct effect of increasing social divisiveness between users[56]. This constant exposure to divisive content is, indeed, resulting in more loneliness and isolation and is leading internet users to find comfort only in the virtual acquaintances that share their ideas; the "disdain of the whole texture of reality […] and isolation and lack of normal social relationships" are, according to Hannah Arendt, some of the main conditions that have given way to totalitarian regimes in the past[57]. By offering "[…] the illusion of companionship without friendship" social media is effectively nudging us towards a world where we won't feel empathy for members of our own species, just like we don't while hiding behind a screen[58] .

Considering all these issues, the increasing presence of the internet and social media in our everyday lives is potentially alarming. In Italy, for instance, the amount of people who have internet coverage has almost tripled between 2001(27.2%) and 2021(74.9%) while the number of active mobile phones was around 78 million according to data collected two years ago, nearly twenty million more than the number of inhabitants of the country[59]. What does this say about our future? Are we on our way to become unconscious slaves of the algorithms behind social media sites? Perhaps an answer to these questions may be found by taking a deeper look at what lies behind these algorithms: artificial intelligence.

[56] Ibid., 166-167.
[57] Hannah Arendt*, The Origins of Totalitarianism* (London: Penguin Books, 2017) **as cited in** Mark Coeckelbergh, *The Political Philosophy of AI*(Cambridge: Polity, 2022), 114-115.
[58] Sherry Turkle, *Alone Together* (New York: Basic Books, 2011) **as cited in** Mark Coeckelbergh, *The Political Philosophy of AI* (Cambridge: Polity, 2022), Ibid.
[59] "Mobile Communications and Internet In Italy", WorldData.Info, accessed June 7th, 2023, https://www.worlddata.info/europe/italy/telecommunication.php#:~:text=Mobile%20communications%20and%20Internet%20in,average%20of%201.3%20per%20person

2.2 Artificial Intelligence: A Hazier Picture

Artificial intelligence has been an object of fascination for human beings since the dawn of time, with legends from different cultures and traditions presenting AI beings as powerful guardians that defended cities or villages from their enemies; like Talos, a bronze giant from Greek myths, or the Golem from Jewish folklore, who was brought to life from clay with the magical rites of the Kabbalah[60]. Today, of course, AI has escaped the myths it originated from and has become an integral part of our lives; indeed, we always have it by our side during our daily routines, in our pockets or in our bags. I am talking, of course, about the virtual assistants that are integrated in our smartphones, like Apple's Siri.

Siri's history starts in 2007, when members of a non-profit company, SRI International, founded Siri Inc. with the goal of finishing what the DARPA (Defense Advanced Research Project Agency) had started years prior: creating a virtual personal assistant[61].
After acquiring it in 2010, Apple started incorporating the system in their iPhones from 2011 and in their iPads from 2012. Siri can understand and speak a variety of languages and is activated with the vocal command "Hey Siri!" followed by a particular request, that must be included in the following:

a. Navigate directions

b. Schedule events and reminders

c. Search the web

d. Relay information

e. Change settings[62]

To perform these tasks Siri must be connected to the internet; it converts the audio it recorded (containing the request made) into data, which is then sent and processed by Apple servers, who will

[60] Mirko D. Garasic, *Leviatano 4.0: Politica Delle Nuove Tecnologie* (Rome: LUISS University Press, 2022), 66.
[61] Erica Mixon and Colin Steele, "Siri", TechTarget, last updated February 2023, https://www.techtarget.com/searchmobilecomputing/definition/Siri
[62] Ibid.

return an answer from a large database of commonly asked questions[63]. Beyond Siri, we are also starting to see intelligent vacuum cleaners, like Roomba, or Alexa and Echo, which, if connected to the electric systems of the house, can perform a variety of domestic-related tasks like turning off the lights, for example[64]. AI beings are now able to listen to every conversation we have, know where we are, where we are going and are aware of our schedule and appointments; this goes hand in hand with totalitarianisms' tendency to spy on their citizens, and confirms AI's potential role as game changer in world politics[65]. The notion that artificially intelligent beings may pose a danger to our societies, however, is not new to the collective consciousness.

Once again, it is relevant to turn to past human folklore; this time jumping ahead a few thousand years from ancient times to post WWI Europe. The term *robot* was used for the first time by Czech author Karel Čapek in his 1920 theater production *Rossum's Universal Robots*, and comes from the Czech term *robota,* which roughly translates to 'hard work'. In this drama robots end up rebelling against their human masters and take control of society, showing just how much subconsciously terrified we are of our own creations turning against us. It is interesting to see how the perception of artificial beings has shifted throughout history; from good-hearted and brainless protectors to ultra-intelligent and ruthless foes[66]. This change might have been due to the previously unseen horrors of the War that had ended just a few years prior, which truly showed just how much destruction men can cause with their creations (i.e., tanks, machine guns, poison gas and bombs), and "ushered in a modern era of technology and capability"[67].

Two decades after Čapek envisaged a dystopia where man-made creatures took over the world, amid WWII, Isaac Asimov, in his short science-fiction story "Runaround", developed his *Three Laws of Robotics* as a sort of solution to the catastrophic vision of Čapek:

> 1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
>
> 2. A robot must obey the orders given to it by human beings, except where such orders would conflict with the First Law.

[63] Ibid.
[64] Garasic, *Leviatano 4.0,* 67.
[65] Mark Coeckelbergh, *The Political Philosophy of AI* (Cambridge: Polity, 2022), 114.
[66] Garasic, *Leviatano 4.0,* 66-67.
[67] Theo Mayer, "Technology and World War I: Then and Now", WorldWarICentennial, accessed June 9th 2023, https://www.worldwar1centennial.org/index.php/communicate/press-media/wwi-centennial-news/6968-technology-wwi-then-and-now.html

3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law[68].

In Asimov's world these laws are integrated into robots' programs, meaning that they can't be bypassed, and act as a sort of precautionary measure to keep these artificial beings in check. However still, they aren't flawless, and the author was very much aware of that; indeed, in most of his stories, the main challenges robots faced are related, if not directly caused, by the consequences of them following the rules they were given[69].

Eighty years later, we have come to a point where artificially intelligent programs are being used not only to clean our houses or help us in our daily tasks but, also, much to both Asimov and Čapek's horror, to cause harm and destruction. For example, the South Korean military forces use an automatic machine gun, able to identify and eliminate its targets without any human input, while Russia employs intelligent tanks, known as *robocops*, to patrol sensible areas. In 2005 the Pentagon's plans to substitute human soldiers with artificial ones came to light in a *New York Times* article, with the benefits being very clear to see: the loss of human lives would be prevented, and soldiers would not experience PTSD; still, the very fact that human casualties would be left out of the equation could potentially lead heads of state to start seeing war as a more attractive solution to international disputes[70]. Furthermore, the involvement of artificial beings in warfare would raise ethical question regarding the responsibility of their actions and their agency: who should be held accountable for them? The creators of the robots, the military who employed them, the politicians who authorized their usage, or the robot themselves? Should they be treated like full-on moral and legal agents[71]?

The debate on whether AI beings should be made moral agents and/or political agents has been going on for some time now; one of the strongest objections to granting either status to them is linked to their lack of consciousness and capacity for suffering. However, some have argued that robots and AI beings could be given rights not because of their moral status, but because of their political significance; these rights would be assigned to robots based on their exploitative potential (how they could be used to humanity's advantage), something that could be problematic because it would echo

[68] Isaac Asimov, "Runaround", *Astounding Science Fiction*, 3(1942) **as cited in** Jeremy H. Norman, "Asimov's Three Laws of Robotics + The Zeroth Law", HistoryofInformation, accessed June 9th, 2023 https://www.historyofinformation.com/detail.php?entryid=4108
[69] Norman, "Asimov's Three Laws of Robotics".
[70] Garasic, *Leviatano 4.0*, 94-96.
[71] Ibid., 96-97.

the racial colonialist regimes of our past[72]. There is then a grimmer picture, one which is, according to some, already a reality. AI machines are able to "suggest, enable, solicit, prompt, encourage, and prohibit certain actions, thoughts, and affects", particularly when it comes to international politics[73]; this process will most likely lead to a form of enslavement, which, it could be argued, every person is helping to advance. By using technology in our everyday lives, we are funding corporations that are set on "capital accumulation, support specific hegemonic societal structures, reinforce binaries, and deny pluralities": in short, human themselves, according to this view, are bringing about a new era of AI technocracies[74].

Regardless of the doubts, fears, and speculations about the moral and political standing of AI and their potentially catastrophic consequences, at the end of the day, we are still giving money to the corporations which produce these machines; despite the bleak visions of Čapek and Asimov being represented over and over in media, humanity still seems ignorant of the dangers that artificial intelligence may pose to our lives. Big corporations are, indeed, continuously capitalizing on our desires for new AI tech and some magnates are ready to take the next step; that of AI and human integration.

[72] Mark Coeckelbergh, *The Political Philosophy of AI* (Cambridge: Polity, 2022), 180-82.
[73] Maurizio Lazzarato, *Signs and Machines: Capitalism and The Production of Subjectivity* (Los Angeles:Semiotext(e), 2014) **as cited in** Mark Coeckelbergh, *The Political Philosophy of AI*(Cambridge: Polity, 2022), 188.
[74] Mark Coeckelbergh, *The Political Philosophy of AI* (Cambridge: Polity, 2022), Ibid.

2.3 The Next Step: Brain Computer Interfaces

More than forty years ago, Foucault gave a definition of what he thought constituted a human being:

> The individual, with his identity and characteristics, is the product of a relation of power exercised over bodies, multiplicities, movements, desires, forces[75].

We have come to a point in our history, which Foucault would have struggled to imagine, where this definition does not, indeed, apply only to human individuals, but also to other artificially intelligent entities. Years of research have led scientists to connect the human brain to an AI program to perform various tasks, revolutionizing this relation of power.

Scientists and researchers refer to this surprising new development as *Brain Computer Interfaces*, more commonly referred to as BCIs. A BCI can be defined as such:

> A computer-based system that acquires brain signals, analyzes them, and translates them into commands that are relayed to an output device to carry out a desired action[76].

A Brain Computer Interface, thus, shares a lot in common with virtual assistants, such as Siri: both are programs that receive determinate signals, transform them into specific data and send it to an external source to perform a particular task. The key difference between them, of course, is that in one there is a direct connection to a human being, as in BCIs the brain signals are acquired through the recording of specific electrodes which are found either on the brain's cortical surface or on the scalp. The potential of BCIs is limitless, as any person with a disability that prevents them from moving their limbs, or speak, could use them to control robotic prostheses, move a wheelchair or have their needs displayed as written text on a digital screen; however, there are still a lot of issues

[75] Michel Foucault, *Power/ Knowledge* (New York: Vintage Books, 1980) **as cited in** Mark Coeckelbergh, *The Political Philosophy of AI* (Cambridge: Polity, 2022), 146.
[76] Jerry J. Shih, Dean J. Krusienski and Jonathan R. Wolpaw. "Brain-Computer Interfaces in Medicine". *Mayo Clinic Proceedings* 87, no 3 (March 1st, 2012): 268.

relating to the reliability and sophistication of signal acquisition of BCIs, holding them back from being employed more widely on a clinical level[77].

Despite the scientific consensus on BCIs being that they are definitely not ready for qualified medical purposes, some influential and powerful magnates were not about to close the door that those interfaces opened. In 2017, Elon Musk founded *Neuralink* with the goal of defeating neurodegenerative illnesses (like Alzheimer's or epilepsy) with cerebral implants that made use of AI tech[78]. The main issue, however, lies in the fact that the operation necessary to place these implants requires a level of precision that makes it impossible for a human being to perform it; the only way for the procedure to be safely carried out would be if it was done by a robot, specifically designed for this purpose, increasing the costs of the whole project by a large number. Musk, however, predicted that once the operation has been tested, and has been shown to be safe, many people would start wanting these chips implanted in their brains also for recreative purposes (like connecting their thoughts directly to an AI program, such as Alexa, to turn off the lights or to zap TV channels); thus, funding Neuralink procedures and further research, which could help discover new undiscovered areas of the human brain and new ways to cure various diseases[79].

Regardless of how successful *Neuralink* will be in the future, what is undeniably clear to anyone who is observing market fluctuations, is that investors are focusing more and more on this sector, with some estimates claiming that many companies are willing to put in 100 million dollars a year to fund further research aimed at pushing AI-human integration forward. These data confirms that one day what now seems like something out of the pages of Asimov's stories will inevitably become part of our reality: a biotechnologically enhanced human being with limitless potential. Many experts are starting to refer to these future beings as posthumans[80].

[77] Ibid., 270-1; 276.
[78] Garasic, *Leviatano 4.0*, 170.
[79] Ibid., 171-172.
[80] Ibid., 175.

2.4 The Posthuman Question

Before discussing the issues and moral dilemmas that come from the very existence of posthumans further, it is perhaps relevant to shed some clarity on the term *posthuman* itself:

> The "posthuman" is an umbrella term frequently employed in a number of theoretical and critical discourses. It is difficult to find a definition of the term that is shared by all the different approaches that use it, since "posthuman" seems to denote a very diverse group of phenomena, some ongoing and others only predicted or imagined.The "posthuman" is used to describe modes of being resulting from potential enhancements to human nature generated to applied science and technological developments[81].

What the passage cited above makes abundantly clear is that many discussions regarding posthumans are still very much in the speculative territory, if not completely imaginative; indeed, the origins of the term itself can be traced back to science-fiction[82]. However, the goal of this paper is not to argue in favor or against the appropriateness of the term nor on the validity of the speculations surrounding it: it is merely presenting predictions of what these beings may look or act like (and the underlying issues that may stem from this) based on the current direction of technological and biotechnological progress.

The first position worth considering is that of posthumans interpreted as AI/human hybrids, which was already hinted at in section 2.3 and is also, if not the most likely, certainly the most realistic based on current technological trends. Dr. Mirko Garasic argues that there are four main points, or issues, to keep in mind when talking about the future hybridization of humanity. First, there is the issue of how much agency the human being will have and just how much of it will be delegated to the AI within them: should we allow the AI to decide through its algorithm what's best for our lives, from the choice of the better meal to whom we should marry? It has already been shown in experiments how allowing chips in the brain to control our decisions through electrical shocks may prove to be beneficial in different scenarios, but there are still doubts regarding the long-term consequences of this process. Furthermore, the very existence of these hybrids will create new forms of discrimination,

[81] Daniele Rugo, "Posthuman", OxfordResearchEncyclopedias, published on June 13th 2020 https://oxfordre.com/literature/display/10.1093/acrefore/9780190201098.001.0001/acrefore-9780190201098-e-1136
[82] Ibid.

amplify the digital and technological divide between nations and challenge outdated conceptions of what constitutes our egos. Thirdly the development of the technologies required for these hybrids would face the issue of impartiality: humans are by nature biased, they have their own views and prejudices, and that is always reflected in their creations. Finally, Garasic also mentions the issue of privacy as the developers of the AIs might gain unlimited access to the thoughts of millions of people[83].

The second position that will be analyzed in this paper is more complex and slightly more speculative, but nevertheless very relevant, particularly for discussions regarding utility maximization. David Pearce first hints at how posthumans may be conceived, thanks to a biotechnological manipulation of the genome: future parents may be able to select in advance the physical and behavioral traits of their child, effectively producing a "[…] supergenius who grows up to be faster than Usain Bolt, more beautiful than Marilyn Monroe, more saintly than Nelson Mandela, more creative than Shakespeare - and smarter than Einstein[84]". This would technically maximize the utility of both the parents and the child, but Pearce also argues that the development of this particular procedure is moving at a relatively slow pace. Regardless of this, however, Pearce argues that whatever the future may bring, researchers should focus on enhancing the biological side of posthumans and give less power to the AIs: according to him, while AI programs are getting more intelligent, they are very far from being conscious of their own existence, making them "zombies[85]".

Pearce, thus, starts to focus on figuring out how posthuman biology can be enhanced to maximize utility overall. The main issue to address here is that of the inherent competitiveness human beings have against other species and, as time went on, increasingly against their own; just in the twentieth century alone humans killed around 100 million of their kind. The laziest and most feasible solution today is to simply inject humans with drugs, like oxytocin (the so called 'trust hormone'), which could technically be used to contrast the violence provoked by testosterone in human brains: this would translate in less crimes of territorial aggression, but, if one considers that male competitiveness is also the main reason for most technological advancements throughout history, it becomes clear that drug prescription can't possibly be the long-lasting solution we are looking for[86]. Furthermore, this would also face the same issue of Nozick's experience machine argument, which was extensively

[83] Garasic, *Leviatano 4.0*, 175-6.
[84] David Pearce, "The Bio intelligence Explosion: how recursively self-improving organic robots will modify their own source code and bootstrap our way to full-spectrum superintelligence", BioIntelligenceExlposion, accessed June 14th 2023
https://www.biointelligence-explosion.com/

[85] Ibid.
[86] Ibid.

discussed in section 1.3(living as addict would surely not make either me or the people around me happy). Because of this, David Pearce suggests that the right path towards the annulment of humans' biological unfriendliness is to somehow set up posthumans to be born with the relatively rare condition of *mirror-touch synesthesia*[87]:

> Mirror touch synesthesia is a condition that causes a person to feel a sensation of touch when they see someone else being touched. The term "mirror" refers to the idea that a person mirrors the sensations they see when someone else is touched. This means when they see a person touched on the left, they feel the touch on the right. […] Researchers theorized that people with mirror touch synesthesia have enhanced sensations of social and cognitive recognition compared with others[88].

Thus, if posthumans became all mirror touch synesthetes, the issue of unfriendliness and distrust would disappear; as no one could or would want to harm anyone else more than they want to harm themselves, because of their capacity to assume the first-person perspective of the other[89]. An estimate made by the University of Delaware found that, currently, around 2% of the world population has this condition, which has also been shown to be often correlated with autistic symptoms[90]. Indeed, David Pearce envisages posthumans as having an autistic-level super-intelligence, not dissimilar to AI, while simultaneously being able to grasp "[…] all possible first-person perspectives[91]" and act taking them into consideration, thus maximizing collective utility. Even after all these centuries, the pleasure-pain axis hasn't lost its relevance: what truly, intimately, makes us human, is our desire to seek pleasure and abhor pain[92]. Indeed, one of the main conditions why some argue against giving moral status to robots is that they lack the capacity to suffer (section 2.2). To sum up, Pearce's most idealized version of a human being is a super-genius who knows everything that there is to know, and is thus fully informed, but who also possesses an innate sense of compassion and understanding of all people's perspectives and of their desires as if they were his own, thus being able to satisfy his own informed preferences and all other people's preferences at the same time.

---

[87] Pearce, "The Bio intelligence Explosion"
[88] Heidi Moawad, "Is Mirror Touch Synesthesia a Real Thing?", Healthline, published January 13th 2020 https://www.healthline.com/health/mental-health/mirror-touch-synesthesia
[89] Pearce, "The Bio intelligence Explosion".
[90] Moawad, "Mirror Touch Synesthesia"
[91] Pearce, "The Bio intelligence Explosion"
[92] Ibid.

Interestingly, the two predictions of posthumans presented in this section mirror humanity's two fictional notions of artificially intelligent creatures: the benevolent guardian of ancient folklore and the cold and dangerous machine from the war ridden world of twentieth century (section 2.2). Indeed, Garasic focused mainly on the issues that the mere existence of an AI-human hybrid may bring more than the benefits, while Pearce based the entirety of his discussion on solving the issues human beings already face and built a posthuman with those issues in mind. Perhaps this recurring duality seen in human folklore, fiction and scientific speculation shows that, at the end of the day, human beings need a clear line that separates what is good from what is bad; that is why the cases discussed in the previous sections can be classified as moral dilemmas at all.

2.5 The Utilitarian Position

The question that titles this chapter may remain, at least for the time being, unanswerable. Most of the arguments presented while talking about social media, artificial intelligence and, more specifically, posthumans, are very much speculative in nature. Although we may not give a definitive answer to this intriguing query, what we can do is reflect on the stance utilitarians would take if faced with the troubling dilemmas technology presents us with.

In the case of social media, surveys did show their positive effects on the general level of satisfaction of the population, even making marginalized people (elders, socially anxious folks etc..) feel better by helping them connect with people online and build new relationships; however, the main pitfall of social media are the AI algorithms that, by showing divisive content to users, push people away from their real life friends towards their virtual acquaintances, without them realizing(section 2.1). This is a textbook example of Richard Hare's so called *unexperienced preferences*; we are not conscious that are lives are worse off, but still, they are[93]. Of course, I consciously would prefer not to be manipulated by the big corporations that own most social media sites, but since I don't experience or am not conscious about the coercion/manipulation going on, my satisfaction levels remain unchanged, explaining why many people in surveys may answer that their experience on social media is overwhelmingly positive. In this case, the utilitarian position seems to paint quite the black and white picture; most utilitarians, from Bentham to Hutcheson, would agree that social media does not maximize collective utility.

When talking about AI robots, however, utilitarians may not be able to give such a clear, unequivocable sentence. On one hand, robots are starting to be employed in very repetitive manual work, preventing human workers from injuring themselves while simultaneously saving companies money; on the other, they are taking the jobs of countless workers and their ability to provide for their families[94]. Furthermore, their employment in warfare may limit human casualties in the short run, but may also push nations, for this very reason, to resort to war more frequently, leading to more casualties, harm, and destruction in the long run (section 2.3). Perhaps utilitarians would take a page from Asimov's book, suggesting that robots are programmed in a way that always maximizes utility (section 2.3); this, however, could still have problematic implications. Consider again, for instance,

[93] Kymlicka, *Contemporary Political Philosophy*, 17.
[94] Brian Greene, "How to shape the future of AI for the good", Markkula Center For Applied Ethics, published December 16th 2016
https://www.scu.edu/ethics/focus-areas/technology-ethics/resources/social-robots-ai-and-ethics/

the example of a policeman whose only pleasure comes from torturing a dreamer (section 1.4). In this scenario the robot would probably let the torture happen, because the overall utility would be higher than if it hadn't. In a similar fashion, if tasked with ensuring world peace, the robot would probably kill millions of terrorists, criminals, and murderers. These scenarios are not far from *teleological utilitarianism*, where "the right act is defined in terms of maximizing the good, rather than in terms of equal consideration of individuals"[95]. Because AI lacks the capacity for suffering itself (section 2.2), it would not feel empathy, and would thus lean towards always maximizing the overall good, no matter the consequences. Indeed, for teleological utilitarians:

> The goal is not to respect *people*, for whom certain things are needed or wanted, but rather to respect the *good*, to which certain people may or may not be useful contributors[96].

Though arguably free of the contradictions and flaws of other variations of utilitarianism, any human being, utilitarian or not, would surely contest these words as absurd, and so would most of the authors presented in this paper. However still, welfarism (section 1.4) does seem to require at least some level of apathy for its main arguments to work, as seen with the policeman and dreamer example. Unsurprisingly, utilitarians would be divided on the wider employment of AI robots: the teleological utilitarians and welfarists would see in them the ideal utilitarian agent, while Bentham, Cumberland, Mill and Sidgwick would be, if not completely against it, way more cautious about it.

If the question of AI robots split utilitarianism down the middle, then the more speculative discussion on posthumans would surely cause much more problems. Two predictions of how posthumans may turn out to be were discussed in the previous section: Garasic's AI/human hybrid and the biotechnologically enhanced mirror synesthete of Pearce (section 2.4). Garasic, looking at the current direction of investments and projects like BCIs and Musk's Neuralink, observed that the most realistic version of future posthumans would be one were humans became one with machines, so to say. He himself painted quite the negative picture of these beings however, and most utilitarians would agree with him; having a chip implanted in our brains that can read our thoughts and act for us may raise questions about our agency, making our status as moral agents precarious, and of our privacy, as the developers of said chip will gain access to our thoughts. Once again, perhaps, the only ones who

[95] Kymlicka, *Contemporary Political Philosophy*, 33.
[96] Ibid., 36.

would accept Garasic's vision are the teleological utilitarians, who would see the benefits of monitoring and controlling the actions and thoughts of millions, in order to maximize the overall good; while completely neglecting basic human rights, of course. Pearce's alternative, on the other hand, recognizes the dangers of integrating AI with human beings and immediately excludes it from his argument: he envisages a being who, thanks to his mirror touch synesthesia, is able to assume the perspectives of every person as if they were his own, and act taking them into account (section 2.4). In the last section. I briefly mentioned how we can intuitively tell, between these two predictions, which is the better one, which is the good one, just like utilitarians and their precursors would. From Mozi to Sidgwick, from Epicurus to Bentham, all would agree that the mirror synesthete is the perfect moral agent, but perhaps none more than the teleological utilitarians themselves; if everyone is literally incapable of making decisions that negatively affect the happiness of others, then the collective good would be maximized, while at the same time, to the relief of everyone else, it would not threaten the basic dignity of human persons. Indeed, the main attraction of utilitarianism, at its core, is that it "[…] conforms to our intuition that human well-being matters[97]"; once the literature became more complex, utilitarianism lost much of its appeal (section 1.5).

In the end, it all leads back to the pleasure and pain axis. AI cannot suffer, it cannot empathize with human beings and their suffering and is not conscious about it; this makes it alien to us, and thus giving it more and more power over ourselves seems self-defeating (sections 2.2, 2.4). Pearce's solution to this problem is to enhance the human being in a way so that the empathy he already feels for others is not just conditional (based on circumstance) but universal. But is the mirror synesthete the perfect utilitarian agent? Would he act differently than a regular human or a robot when put on the field, when faced with moral dilemmas? The next chapter will hopefully help in answering this question.

[97]Kymlicka, *Contemporary Political Philosophy*, 12.

3. Greene's Trolley Problems

3.1 The Three Scenarios: Switch, Footbridge, Loop

The 21st century saw a growing interest in cognitive science, with experimental psychologist and neuroscientist Joshua Greene arguing that important conclusions can be drawn from how we make moral decisions. To this end, Greene employed three imaginative scenarios meant to challenge our sense of everyday morality, that is, what we intuitively know to be inherently right or wrong.
The problems or scenarios are the following:

- ***Switch:*** […] a runaway trolley is heading down a railway track. If you do nothing, it will kill five people. The only thing you can do to save the five is pull a switch that will divert it down a sidetrack, where it will kill only one person. […] (All the people are strangers, and you know no details about them).

- ***Footbridge:*** […] there is again a runaway trolley that will kill five people unless you act, but this time you are standing on a footbridge over the tracks and there is no switch. You think about sacrificing your own life by jumping onto the track in front of the trolley, but you realize you are too light to stop it. A stranger wearing a heavy backpack is standing next to you, however, leaning over the rail. The only thing you can do to save the five is push him off the footbridge onto the track in front of the trolley. He will be killed, but the weight of his backpack will stop the trolley before it hits the five.

- ***Loop:*** […] you can pull a switch to divert the trolley, but this time the side-track loops back onto the main track, where it would still kill the five, were it not for a stranger who is asleep across the tracks. The trolley will hit that person, killing him, but his body will stop it going any further, so the five will live[98].

[98] Radek and Singer, *Utilitarianism*, 32-34.

3.2 A Conditional Aversion to Violence: The Human Reaction

In previous sections of this paper, I suggested that analyzing certain utilitarian scenarios (and humans' reaction to them) could prove useful in trying to grasp what it is that makes us human, what separates us from machines. In the last chapter, despite Pearce's remark about the natural unfriendliness humans have towards one another (section 2.4), an apparent answer to this dilemma was reached. Our capacity for suffering, our empathy, our understanding of other people's first-person perspectives is what makes human. The main issue with this, something that Pearce recognized and set out to address in his research, is the conditionality of said empathy (sections 2.4, 2.5): if the circumstances we find ourselves in change, even slightly, our decisions may shift drastically. Nothing can exemplify this better than Greene's trolley problems.

The average reaction to the problems presented in section 3.1 was interesting; both in *Loop* and in *Switch* most people believed that they would be capable to kill one person to save the five, but not in *Footbridge*. To understand why people reacted like this, Greene asked some respondents to imagine these scenarios while undergoing magnetic resonance: he found that in *Switch* (and *Loop*) the areas of the brain related to cognition were more active, while in *Footbridge* he observed more activity in the side of the brain related to emotion[99]. These data contributed greatly to the formulation of the *dual process theory* of moral reasoning, which Greene explained by comparing it to a camera:

> […] (A camera) has an automatic 'point and shoot' mode, as well as a manual mode. For taking photographs in everyday situations, 'point and shoot' is quick, convenient, and generally gives better results than people with limited time and no special expertise would get by using manual mode. In special circumstances, however, when the light is unusual, or we are trying to achieve a particular effect, we will do better to adjust the settings ourselves, taking time to work out what will give us the best result[100].

The *Switch* and *Loop* scenarios are perfect examples of the "special circumstances" mentioned above, where logic triumphs over emotion. According to Greene, the main reason why many people said they could not bring themselves to kill the stranger in *Footbridge* is because of the intuition that we

[99] Radek and Singer, *Utilitarianism*, 34-35.
[100] Ibid., 35-36.

all have, since ancient times, that human on human physical violence is always wrong[101]; because of this, when faced with problems such as those described here, humans revert to their default "point and shoot" mode, taking the easy way out. However, if we were to take out the variable of direct violence out of the equation, would our reaction remain the same? To answer this question, Greene created a variation of *Footbridge*, *Remote Footbridge*:

- ***Remote Footbridge***: Once again there is a runaway trolley and a stranger on a footbridge, but this time you are not on the footbridge. Instead, you are standing next to a switch that will open a trapdoor over which the stranger is standing, causing him to fall onto the track and be killed, but saving the five[102].

The reaction to this version of the trolley problem was very different from the original, with around 63% of respondents allowing the stranger to be killed in *Remote* compared to 31% in *Footbridge*. This shows how people have different reactions based on factors that are completely irrelevant, at least on utilitarian grounds. Greene argues that our intuitions of what is right or wrong, though at times vulnerable to irrelevant elements, have become so universal because they have been tested over millennia, and proven to yield good; thus, it would not be wise to dismiss those that have reactions based on these intuitions[103].

In the first chapter of this paper, while talking about the flaws of welfarism, I briefly brought up Richard Hare and his 'level 1'/ 'level 2' thinking, which mirrors almost perfectly Greene's *dual process theory*. In both theories we have two dimensions of moral reasoning; one to be used in 'point and shoot'/ 'fight or flee' situations, based on intuition, and another to be used in more complex, "special" scenarios, based on critical thinking (section 1.4). Just like Greene, Hare does not dismiss our moral intuitions: on the contrary, he believes that we should follow them in our everyday lives as much as possible. However, he argues that there are situations where the picture is hazier, more complex, and following our instincts may not yield the best outcome. In these scenarios, echoing Greene, we need to adjust ourselves the settings of our camera, so to say, and evaluate the consequences of our actions taking into consideration "[…]all morally relevant properties of the

[101] Radek and Singer, *Utilitarianism*, 35.
[102] Ibid., 36.
[103] Ibid., 37.

choice at issue"; this implies that while taking a stance on a certain matter, we need to take into consideration all those affected by the decision because, once we have, it will be valid "[…] in both hypothetical and real cases[104]". In the trolley problems, for instance, we may come to the conclusion that it is permissible to kill one person by activating a switch; but if we do, than we must also concede that this decision would stand even if we had to push them ourselves on the rails, if it was one of our relatives to die, or even if it was ourselves that would be sacrificed in that particular scenario. In fact:

> The critical thinking proposed by Hare thus requires us to engage in role taking, putting ourselves in the position of others: considering the effects of a choice on everyone and aggregating these effects on everyone's preferences as if they were conflicting preferences within the thinker herself[105].

Hare's theory does follow a welfarist logic, in part to its detriment, but it also stresses how important it is that we understand the gravity of certain judgements we make by adopting a criterion of universality[106], something that directly links his arguments with Pearce's posthuman.

Despite all the theory and literature available, however, data showed how humans themselves are not perfect, and that "[…] universal judgments made at the critical level can conflict with intuitive judgments[107]". Our intuitions are so deeply rooted in our psyche that we may not be able to consciously recognize when it would be appropriate or not appropriate to follow them. Humans are creatures that have relationships, commitments, beliefs, emotions, and mental states and all are unique in their own ways: generalizing any behavioral pattern to humanity is simply impossible. This, however, is not true for robots.

[104] Jonathan Baron, "Richard M. Hare", Utilitarianism, accessed July 8th 2023 https://utilitarianism.net/utilitarian-thinker/richard-hare/
[105] Ibid.
[106] Ibid.
[107] Ibid.

3.3 Torn Between Utilitarianism and Asimov's Laws: The Robots' Reaction

The dangers of artificial intelligence were discussed at length in the previous chapters of this paper (sections 2.2;2.5) so I won't dwell on them further. Instead, it would be interesting to speculate on how robots would behave when faced with the trolley problems. First, however, we would need to clarify what kind of robots we would be dealing with. For the purposes of this discussion, they would need to have limbs, in order to be able to push people on the railway and be completely independent from human input. Secondly, we would have to deal with the question of their programming; technology itself, as Garasic argues[108], is impartial, but its employment by human beings may yield either positive or negative consequences. That is why Garasic suggests that laws should be put in place to regulate the employment of these technologies; in the case of robots, for instance, he briefly mentions Asimov's *Three Laws of Robotics* (section 2.2) as a potential tool to make sure these artificially intelligent creatures don't become a threat to humanity[109]. For the purposes of this paper, both Asimov's laws and utilitarian logic will be considered as a basis for the robots' programs. Furthermore, before discussing the robots' reaction to the trolley problems, it is important to clarify that the *dual process theory* and Hare's *levels theory* will not apply to robots, as they don't have intuitions, they are only capable of thinking in a critical way, following a series of algorithms, in a manner not dissimilar from virtual assistants, such as Siri (section 2.2).

The first thing to note is that robots are effectively emotionless, they don't feel empathy, they can't, in Hare's words, "engage in role taking" and thus the important variable of physical violence, discussed in the previous section, would be treated just like the irrelevant element it is in utility calculations; the robots would see the three(four scenarios, if one counts *Remote Footbridge*) scenarios as practically identical. They have a choice to either let one person, or five, die. Now, any human would surely concede that saving five lives instead of one would be better, but this is because of another variable of the trolley problems: we don't know these people, we don't know their past or their criminal record, they are strangers to us. This would not be a problem for the robots. Face recognition algorithms have started being employed in smartphones for some time now, particularly in iPhones, where a special infrared camera takes a picture of a person's face and transforms it into a numerical code to identify them[110]; it would not be far-fetched, then, to assume that robots tasked with facing such dilemmas would possess this feature. They would be able to identify these strangers and know everything about them, and thus be "fully informed" (section 1.3). Echoing the *Baby Hitler*

[108] Garasic, *Leviatano 4.0,* 186-187.
[109] Ibid., 68-69.
[110] Dreamchild Obari, "What Is Apple's Face ID and How Does It Work?", MakeUseOf, published June 12th, 2023.

example (section 1.4), the robot may figure out that, out of the five strangers, three are murderers, or are likely to commit a series of murders soon, based on their history, family life or mental health. In this scenario, the robot would probably let the trolley hit the five and would simply consider the two innocent men to be collateral damage, in a very teleological fashion (section 2.5).

But what if the robots where also programmed to abide by Asimov's laws? This scenario could potentially turn out to be even more grim. The first and most important law literally prohibits robots from doing anything that could harm a human being, even because of their "inaction"; in addition to that, robots are also expected to sacrifice themselves if that would help save human lives, as made clear in the third law (section 2.2). This would leave them in a very difficult spot: they may be aware of how dangerous some of the individuals on the rails are or will/may become, but their program would not be able to let them die, because of the inherent value of human life. This would most likely result in the robot sacrificing itself to save all six people but, realistically, some, if not all of them, would perish anyway. Asimov's laws do help in creating internal conflicts that make for interesting reads, but if applied to reality, their consequences would be dire to say the least.

If we were faced with the prospect of either Asimov's or a utilitarian robot, we would intuitively lean more towards the latter; yes, its decisions may be cold and ruthless, but at least we would have the certainty that some of the six people on the rails would make it out alive. However still, these utilitarian robots, with their teleological tendencies, are far from an ideal moral agent and most utilitarians themselves would agree. With the danger of robots treating people just as location of utilities (section 2.5), and the everlasting conflict between intuitive and critical thought burdening human beings, would the posthumans fare any better?

3.4 A Speculative Mystery: The Posthuman Reaction

All discussions regarding the suitability of human beings and AI robots as the perfect moral agent were essentially trying to find an answer to a rhetorical question. Both are inadequate for different reasons, but they do individually have qualities which could, if one being somehow possessed them at once, lead to solving the dilemma behind the *dual process theory*. As explained in section 3.2, human beings possess conscience and critical thinking; the problem lies in the fact that, as shown in *Footbridge*, humans also act on the basis of their intuitions, which are so strong that they often trump any rational or critical judgment they may have made up in their minds. The solution, as Pearce noted (section 2.4), can't possibly lie in integrating humans with AI robots, because they are not conscious about the pain they may inflict with their actions; indeed, they are able to only think critically, as they see persons just as numerical variables in a mathematical problem their algorithm needs to solve. Pearce's posthuman was practically crafted from scratch considering all the issues human beings face when it comes to them being completely morally righteous, specifically in a utilitarian sense. Would this be enough to pass this final test? Would the mirror synesthete turn out to be the ideal utilitarian agent Pearce made it out to be?

Before directly bringing in the various trolley cases, it is important to remind ourselves of the qualities Pearce's mirror synesthete possesses, in his words:

> We're not all closet utilitarians. Genghis Khan wasn't trying to spread universal bliss. As Plato observed, "Pleasure is the greatest incentive to evil." But here's the critical point. Full-spectrum superintelligence entails the cognitive capacity impartially to grasp all possible first-person perspectives - overcoming egocentric, anthropocentric, and ethnocentric bias (*cf.* mirror-touch synaesthesia). As an idealisation, at least, full-spectrum superintelligence understands and weighs the full range of first-person facts. [….] If your hand is in the fire, you reflexively withdraw it. In withdrawing your hand, there is no question of first attempting to solve the Is-Ought problem in meta-ethics and trying logically to derive an "ought" from an "is". Normativity is built into the nature of the aversive experience itself: I-ought-not-to-be-in-this-dreadful-state. By extension, perhaps a full-spectrum superintelligence will perform cosmic felicific calculus and execute some sort of metaphorical hand-withdrawal for all accessible suffering sentience in its forward light-cone. Indeed, one possible *criterion* of

> full-spectrum superintelligence is the propagation of subjectively hypervaluable states on a cosmological scale[111].

On the one hand, this posthuman seems to perfectly go along with Hare's theory, and certainly would be able to put himself in the shoes of anyone more than the average person; but perhaps that's where the main problem lies. Pearce seems to suggest that the posthuman would be able to immediately solve any problem intuitively, without any critical thinking to be had(as seen with the hand in the fire example), but on a universal scale, as his condition would make it so that he "must ultimately do what a classical utilitarian ethic dictates and propagate some kind of "utilitronium shockwave" across the cosmos[112]", because if he'd truly wants to maximize his own utility, he would need to maximize all other persons' utilities at once. Paradoxically, the most recent theory analyzed in this paper seems to return to Cumberland's theological utilitarianism; the mirror synesthete resembles more the Abrahamic God than anything even remotely close to a human being. Here is a description of the Christian God, from the Bible:

> O the Lord, the Lord God, merciful and gracious, patient and of much compassion, and true, who keepest mercy unto thousands: who takest away iniquity, and wickedness, and sin, and no man of himself is innocent before thee. Who renderest the iniquity of the fathers to the children, and to the grandchildren unto the third and fourth generation[113].

Pearce's posthuman theory basically reads like an updated version of this quote from the book of Exodus, and even he himself acknowledges this;

> Man proverbially created God in his own image. In the age of the digital computer, humans conceive God-like *super*intelligence in the image of our dominant technology and personal cognitive style - refracted, distorted and extrapolated for sure, but still through the lens of human concepts[114].

[111] Pearce, "The Bio intelligence Explosion"
[112] Ibid.
[113] Ex. 36: 6-7 DRB
[114] Pearce, "The Bio intelligence Explosion"

Perhaps the main difference between God and the posthuman is that the latter technically has a human body, biotechnologically enhanced to be smarter, better looking and more compassionate, yes (section 2.4), but still limited in its agency; he cannot part the seas or send down plagues from the heavens. So how, exactly, would this utilitarian demigod act when faced with the dilemma of the trolley problems?

The first thing to consider is that the mirror synesthete would feel all the pain of anyone being hit by the trolley as if it was himself being hit, so he would act in a way so that the least amount of people would be hit; but would he be able to cause the death of even a single person and endure the pain that would result from it? We know that this posthuman has "a metric to distinguish the important from the trivial" and an "autistic, pattern-matching, rule-following, mathematico-linguistic intelligence[115]", so he would surely know what the best course of action would be, in his mind at least, but he would have to have the strength to actually go through with it, enduring incommensurable pain for the greater good, in this case, for the maximization of his and others' utilities. In *Switch*, *Footbridge*, *Loop* and *Remote Footbridge* he would likely understand that sacrificing one person would be better than killing five; however, unlike the robots, he would not know anything about them apart from what they are feeling in that particular moment. So, what would he do? Would he be overwhelmed with emotion like the humans were in *Footbridge*, or would he make the ultimate sacrifice to teleologically maximize the overall utility, like the robots would? The truth is, simply, that we don't know.

All the speculation, all the hundreds of years of literature that were quoted, discussed, and mentioned in this paper, eventually brought us back to human myth and folklore; the posthuman, the 'ideal' moral agent, on paper, is nothing more than a new-age, hyper-anthropocentric version of the Abrahamic God. It seems like, in trying to create the perfect utilitarian agent, Pearce effectively turned back time, and returned to utilitarianism's precursors, who saw in something bigger than themselves the justification for their claims. But where does that leave us?

[115] Pearce, "The Bio intelligence Explosion".

Conclusion

This paper's goal was to use utilitarian arguments to challenge our conception of what is a moral agent, and if other beings other than regular human beings could be considered as such. To this purpose, hundreds of years of utilitarian literature were analyzed, and a dozen authors' convictions discussed. Sometimes, technology proved to be a hindrance to some utilitarians' arguments, like in the hedonists' case (Nozick's experience machine); but in others, it was human beings themselves who 'posed a threat' to utilitarian logic, like in welfarist or in preference satisfaction scenarios. For those argumentations to work, the moral agent would need to completely disregard their commitments, personal relationships, and, most of all, their intuitions. The biggest attraction utilitarianism has is that it tickles our intuition that human well-being matters, but, as seen in this paper, rarely fully commits to it. For instance, a welfarist may concede that someone can be tortured, if it maximizes collective utility; just like a preference utilitarian may push a religious man to not go to church, and instead go enjoy a day off in the countryside, even if it would make him miserable (sections 1.3, 1.4). The truth is, simply, that utilitarianism cannot work properly if the agent is human.

With this in mind, the natural question this paper asked was: but what if the agent was not human? Would utilitarian argumentations work then? The answer, as seen in previous chapters, is complex and multifaceted. If the agent was an artificial intelligence, it could be simply programmed to abide by utilitarian arguments and so it would become the perfect utilitarian agent, but it would also risk falling in a teleological pit: treating human beings as contingent variables, in a utility maximization problem on a universal scale (section 2.5). Utilitarianism, in this scenario, would work theoretically, but it would also go against its strongest attraction, making it self-defeating. Who could trust a robot with one's life if it would not hesitate to kill us to maximize collective utility?

Because of this, Pearce distanced himself from AI and sought to create an idealized version of human beings, a posthuman, who could not be able to cause pain to others more than he would want to cause pain to himself, because of his condition as a mirror synesthete (section 2.4). But while artificially intelligent robots already exist today, the speculative nature of Pearce's posthuman made it closer to a mythological figure, a demigod, than a natural evolution of a human being. In the end, as briefly mentioned in section 3.4, Pearce effectively circled back to Cumberland: when utilitarianism exhausts its arguments, it looks as something greater than itself, it leans more into commonly shared beliefs and intuitions. In Cumberland's time, religion seemed to go hand in hand with early utilitarians' message, so he used Christianity to make his theory more appealing. Pearce, similarly, as he himself admits (section 3.4), created a new type of 'deity' who conforms to our 21st century expectation of

what a God should look, think, and act like. The dangers of artificial intelligence gaining more and more prominence in our world nudged certain theorists towards arguing against it, and with good reason (section 2.2, 2.5). However, the arguments that they have come forward with, like in Pearce's case, rely a lot on myth and fiction; in fact, Pearce's posthuman would not be out of place in one of Čapek's theater productions, or in one of Asimov's stories (section 2.2). Of course, Pearce's argument was not based merely on his imagination, but the actual scientific bases he adopted were so experimental that they might as well have been. Until the posthuman escapes the shaky grounds of myths and scientific speculation, like the robots did, we can't possibly consider it as a viable moral agent, just like we can't with God, the mythical Greek bronze giant of Talos, or the Golems from Jewish folklore (sections 2.2, 3.4).

With robots being potentially the catalyst for the doom of our species, given their tendency to look at the bigger picture and fail to understand human pain, and the posthumans being light years away from our collective consciousness, we only have ourselves. We are fallible, imperfect and can certainly be irrational, but when it comes to promoting the welfare of our own species, we may be the only true potential candidates.

In the introduction of this paper, I introduced the concept of our intuitions, whose importance was recognized by some of the theorists hereby analyzed, specifically by Greene and Hare. Both Greene and Hare specified how human beings should be able to recognize when it would be appropriate to listen to their intuitions and when not to, but the problem lied in the fact that most people simply can't do so and will end up always following their instincts. Does this make human beings not suitable as moral agents?

Many of the scenarios that were presented in this paper expect the reader to have a very strong and immediate reaction due to their intuitions, to use their 'point and shoot mode' (section 3.2); when seeing a drowning child or a person being tortured(section 1.4), not a single person, not even a utilitarian thinker, would stop and reflect on what the best course of action would be, they would simply intervene, regardless of possible future consequences of said act. Yes, these consequences may be tragic, but they are completely outside of the agent's influence. If we were to assume that all the consequences of a human act were attributable to the person who committed that act, regardless of time and scope, then we would all either be serial killers or saints.
For instance, a woman claimed that one time she had a terrible feeling about going on a vacation with a group of people, so much so that they had to change plans and postpone the vacation; later, they

found out the plane they were supposed to take crashed[116]. Does that make this woman a seer or an unsung hero? No, just like saving a child that would later grow up to be a genocidal maniac does not make someone a criminal (section 1.4). We may not ever be able to be always fully informed of the facts, and thus be able to express a valid, critical, judgment regarding the situation at hand, but that is simply the reality of being human.

Morality is a multifaceted and complex phenomenon, something that is reflected in the term itself, which can be used:

- descriptively to refer to certain codes of conduct put forward by a society or a group (such as a religion), or accepted by an individual for her own behavior, or
- normatively to refer to a code of conduct that, given specified conditions, would be put forward by all rational people[117].

Utilitarianism seeks to be interpreted normatively as a moral theory, it wants its arguments to be so strong that any rational person would naturally be drawn to them; however, its intricacies, contradictions, and counter-intuitive stances on certain matters, hold it back. However, utilitarianism did serve its purpose in showing how non-human or potentially, posthuman, agents would play against these issues. The main problem, when it comes to arguing in favor of the suitability of a being other than a human having moral agency, as was extensively discussed in this paper, is that they can, indeed, be given a way of discerning what is right from what is wrong, but what seems to escape both robots and posthumans is the delicate balance between empathy and moral righteousness. In the case of robots, they simply can't feel empathy, they can't put themselves in the shoes of anyone, while Pearce's posthuman has the opposite problem; he feels so much empathy towards others that he may not be able to take decisions that involve hurting others, even if that would mean condemning more people to certain death, like in the trolley scenarios. There are certain situations in which our empathic side takes over, and others where we know we must control it to either save ourselves, our friends, or our family members, regardless of utility calculations. Of course, as mentioned while talking about

[116] Alex Gurley, "These People Trusted Their Intuitions and It Saved Their Lives", BuzzFeed, published on December 18th, 2020
https://www.buzzfeed.com/alexgurley/people-share-how-intuition-saved-them
[117] "The Definition of Morality", Stanford Encyclopedia Of Philosophy, last updated September 8th 2020
https://plato.stanford.edu/entries/morality-definition/

the trolley problems, we may prioritize irrelevant elements in our moral decision making, but how irrelevant can they be if a wide number of persons acts in the same way (section 3.2)?

It will take a very, very long time before human beings' moral standing could be seriously put into question, and even then, viable alternatives would have to be considered. This thesis demonstrated that, though humans are not ideal utilitarian agents, they are the only ones who can fully understand the reality of being human, and, because of that, the full spectrum of morality, which is and will remain (at least for the near future) a uniquely human construct. What is undeniable, however, is that artificially intelligent machines' existence and the power they already exercise (either directly or indirectly) in the resolution of international disputes (section 2.2) must be acknowledged and regulated by world powers with appropriate legislation, to avoid the potential catastrophic consequences discussed by Asimov and Čapek. Perhaps in a thousand years, if the world will be taken over by artificial intelligence, no one would even bother with learning about outdated human concepts such as morality or utility; only in that scenario can I imagine the exclusive property of moral agency being taken away from human beings, as their very existence would only serve to further the robot's interests. Until that day comes however, if it ever does indeed, humanity's well-being will depend exclusively on them, as it always has been.

Bibliography

Arendt, Hannah. *The Origins of Totalitarianism* (London: Penguin Books, 2017) **as cited in** Coeckelbergh, Mark. *The Political Philosophy of AI* (Cambridge: Polity, 2022).

Asimov, Isaac. "Runaround". *Astounding Science Fiction*, 3(1942) **as cited in** Norman, H. Jeremy. HistoryofInformation. "Asimov's Three Laws of Robotics + The Zeroth Law". Accessed June 9th, 2023
https://www.historyofinformation.com/detail.php?entryid=4108

Baron, Jonathan. Utilitarianism. "Richard M. Hare". Accessed July 8th, 2023.
https://utilitarianism.net/utilitarian-thinker/richard-hare/

Coeckelbergh, Mark. *The Political Philosophy of AI* (Cambridge: Polity, 2022).

De Lazari-Radek, Katarzyna & Singer, Peter. *Utilitarianism: A Very Short Introduction* (Oxford: Oxford University Press, 2017).

Donatelli, Piergiorgio. *Le storie dell'etica: Tradizioni e problemi* (Roma: Carrocci Editore, 2022).

Foucault, Michel. *Power/ Knowledge* (New York: Vintage Books, 1980) **as cited in** Coeckelbergh, Mark. *The Political Philosophy of AI* (Cambridge: Polity, 2022).

Garasic, D. Mirko. *, Leviatano 4.0: Politica Delle Nuove Tecnologie* (Rome: LUISS University Press, 2022).

Greene, Brian. Markkula Center For Applied Ethics. "How to shape the future of AI for the good". Published December 16th, 2016
https://www.scu.edu/ethics/focus-areas/technology-ethics/resources/social-robots-ai-and-ethics/

Gurley, Alex. BuzzFeed. "These People Trusted Their Intuitions and It Saved Their Lives" published on December 18th, 2020
https://www.buzzfeed.com/alexgurley/people-share-how-intuition-saved-them

Kymlicka, Will. *Contemporary Political Philosophy: An Introduction* (Oxford: Oxford University Press,2002).

Lazzarato, Maurizio. *Signs And Machines: Capitalism and The Production of Subjectivity* (Los Angeles: Semiotext(e), 2014) **as cited in** Coeckelbergh, Mark. *The Political Philosophy of AI* (Cambridge: Polity, 2022).

Mayer, Theo. WorldWorldICentennial. "Technology And WWI: Then and Now". Accessed June 9th 2023, https://www.worldwar1centennial.org/index.php/communicate/press-media/wwi-centennial-news/6968-technology-wwi-then-and-now.html

Mixon, Erica & Steele, Colin. TechTarget. "Siri". Last updated February 2023, https://www.techtarget.com/searchmobilecomputing/definition/Siri

Moawad, Heidi. Healthline. "Is Mirror Touch Synesthesia a Real Thing?". Published January 13th 2020 https://www.healthline.com/health/mental-health/mirror-touch-synesthesia

Mochón, Francisco. "Happiness And Technology: Special Consideration of Digital Technology And Internet". *International Journal of Interactive Multimedia and Artificial Intelligence*, 5 no 3(Dec., 2018), 162-68.

Norman, H. Jeremy. HistoryofInformation. "Asimov's Three Laws of Robotics + The Zeroth Law". Accessed June 9th, 2023
https://www.historyofinformation.com/detail.php?entryid=4108

Obari, Dreamchild. MakeUseOf. "What Is Apple Face ID And How Does It Work?". Published June 12th, 2023.
https://www.makeuseof.com/apple-face-id-explained/?utm_source=flipboard&utm_content=stlevitt%2Fmagazine%2FMac+and+IOS+tips

Pearce, David. BioIntelligenceExlposion. "The Bio intelligence Explosion: how recursively self-improving organic robots will modify their own source code and bootstrap our way to full-spectrum superintelligence". Accessed June 14th, 2023.
https://www.biointelligence-explosion.com/

PhyscologyToday. "Intuitions". Accessed July 18th, 2023.

https://www.psychologytoday.com/us/basics/intuition

Rudolph, Jared. "Consequences And Limits: A Critique of Consequentialism". *Malcaster Journal of Philosophy,* 17 no 1(Jan., 2011): 64-76.

Rugo, Daniele. OxfordResearchEncyclopedias, "Posthuman". Published on June 13th 2020 https://oxfordre.com/literature/display/10.1093/acrefore/9780190201098.001.0001/acrefore-9780190201098-e-1136

Sen, Amartya. "Utilitarianism and Welfarism". *The Journal of Philosophy*, 76 no 9(Sep., 1979): 463-489.

Sher, George. "Justifying Reverse Discrimination in Employment". *Philosophy And Public Affairs*, 4 no 2(1975**) as cited in** Kymlicka, Will. *Contemporary Political Philosophy: An Introduction* (Oxford: Oxford University Press,2002).

Shih, J. Jerry, Krusienski, J. Dean and Wolpaw, R. Jonathan. "Brain-Computer Interfaces in Medicine". *Mayo Clinic Proceedings* 87, no 3 (March 1st, 2012): 268-79.

Smart, C. C. J. "Extreme and Restricted Utilitarianism". *The Philosophical Quarterly*, 6(1956) **as cited in** Rudolph, Jared. "Consequences And Limits: A Critique of Consequentialism". *Malcaster Journal Of Philosophy,* 17 no 1(Jan., 2011): 64-76.

Shaver, Robert. "The Appeal of Utilitarianism". *Utilitas*, 16 No 3(Nov., 2004): 235-250.

The Stanford Encyclopedia Of Philosophy. "Hedonism". Last updated October 17th, 2013, https://plato.stanford.edu/entries/hedonism/#EthHed

The Standford Encyclopedia of Philosophy. "The Definition of Morality". Last updated September 8th, 2020
https://plato.stanford.edu/entries/morality-definition/

Thomson, Judith Jarvis. “Goodness And Utilitarianism”. *Proceedings And Addresses of The American Philosophical Association*, 67, No 4(Jan., 1994) **as cited in** Shaver, Robert. “The Appeal of Utilitarianism”. *Utilitas,* 16, No 3(Nov.,2004): 235-250.

Turkle, Sherry. *Alone Together* (New York: Basic Books, 2011) **as cited in** Coeckelbergh, Mark. *The Political Philosophy of AI* (Cambridge: Polity, 2022).

WorldDataInfo. “Mobile Communications and Internet in Italy”. Accessed June 7th, 2023. https://www.worlddata.info/europe/italy/telecommunication.php#:~:text=Mobile%20communications%20and%20Internet%20in,average%20of%201.3%20per%20person