Ethics for an Open Society

by

José Luis Ferreira

Ethics for an Open Society

by José Luis Ferreira

2025

Ethics International Press, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (Hardback): 978-1-83711-337-8
ISBN (Ebook): 978-1-83711-338-5

Á miña nai, miña naiciña

Table of Contents

About this book

This book has been developing over several years of discussions in class, on social media, and in academic circles. Although demonstrating an objective morality is impossible, the implications of using human moral preferences as a foundation for orderly discussions about ethics are often unclear in moral philosophy, both inside and outside academic settings. This ambiguity may stem from a fear of slipping into unjustified moral relativism. This book aims to bring clarity to moral reasoning based on this foundation. To achieve this, I propose using the tools of modern economic analysis and game theory, which are particularly useful for studying decision-making. The first chapters are dedicated to describing this approach and its justification. Following this, seamlessly, its application to all kinds of moral dilemmas is presented. Although the applications are organized into thematic chapters, this separation is far from rigid, meaning many sections could fit into a different chapter. In some cases, data may be missing, and the conclusions—when they exist—may change upon incorporating new information. A few sections contain mathematical content. If they are difficult to follow, they can be skipped, as they are always accompanied by intuitive explanations that should suffice for a first reading. Originally written in Spanish, the book was translated into English with the assistance of DeepL Translate and ChatGPT. Multiple revisions were occasionally necessary to address misinterpretations.

Acknowledgements

Financial support to write this book by the Spanish Ministerio de Ciencia, Educación y Universidades, through the Agencia Estatal de Investigación project number PID2023-149885OB-I00, is gratefully acknowledged. Some colleagues, friends and family have read earlier versions of the book and have provided valuable insights and comments. Antonio Cabrales, Juan Ignacio Conde and Luis Alfonso Gámez believed in the project and helped in convincing editors and colleagues the book is worth reading. Javier Ruiz-Castillo called me to spend hours discussing the manuscript exhaustively. My sister Pilar Ferreira found many obscure paragraphs, grammatical inconsistencies and typos that needed revision. My friend and colleague Carmelo Núñez did the same and added some suggestions. Gabrielle Gazzei helped with the proofreading. Antonio Gaitán invited me to present parts of the book in his classes on ethics, which helped improving the presentation. My PPE students also helped with their comments when I included some contents of the book in my classes. The blogs *Nada es Gratis* and *Mapping Ignorance* allowed me to reach a wide public, and the numerous comments provided invaluable feedback for the topics in the book. I also must thank the many friends and colleagues with whom I have discussed these topics in the last years. In addition to the above mentioned, philosophers Alicia García, Luc Bovens and Jesús Zamora are some of them. Last, but above all, my deepest gratitude goes to my parents: José Ferreira, a mechanic at Altos Hornos de Vizcaya S.A., and Pilar García. I have never embraced an idea in economic policy—or in life—that I could not explain while looking into the honest, tireless eyes of those who taught me the value of work and dignity.

Manifesto against moral monopolies

There is no original sin, divine mandate or categorical imperative. Nothing is sacred. There has never been a golden age, nor will there be an end of history as long as there are human beings. Essences don't exist, nor does the meaning of life or a purpose in evolution. Nothing is written. Nature is neither good nor bad; it's simply indifferent. There is no natural law. Good and evil don't exist without a moral subject. Society is only a moral subject in a figurative sense. What *ought to be* cannot be deduced from what *is*. It is impossible to discuss morality without considering consequences. The universe doesn't offer a point of view. Time has no original position.

What does exist are human beings making decisions based on their moral preferences, their ability to persuade or impose on one another, and various constraints—whether physical, biological, social, economic, or based on incentives. Good and evil always refer to a being capable of making moral judgments. While one person's morals may not differ drastically from another's, when they do, one's judgment doesn't bind the other. A majority's evaluation doesn't create objective good or evil, not even for that majority. The statement "this is bad" is meaningless without the agreement of those asserting it. The definition agreed upon by a group doesn't bind those who don't share it. One can only say, "I consider this to be bad." Even if something seems particularly bad to me, I cannot simply declare "this is bad," nor can someone else claim to do so with objective certainty. Acknowledging that one cannot say "this is bad" doesn't imply acceptance or resignation; it merely recognizes that moral statements require a grammatical subject that is also a moral subject. The grammatical subject gives the sentence structure, while the moral subject gives morality meaning. An abstract, vague statement strips that meaning away.

Any attempt to deduce a single morality—whether through religious, metaphysical, or ideological conviction—and to impose it—whether through brute force, indoctrination, a majority vote, or militant activism—is an act of intolerance. While one can be intolerant of especially harmful ideas, one should not be intolerant of most moral views that differ from one's own. There can be no moral monopolies.

To impose one individual's morality on another, brute force is sufficient. But for persuasion or agreement among individuals, some ideas are more effective than others. Understanding and studying those ideas is worthwhile. That would make for a good book on ethics. Even ideas relating to the nonexistent can be useful if understood as metaphors or provisional starting points, to be reevaluated as more convincing and useful alternatives are found. I may have slipped when I said "it's preferable" when I should have said "I prefer," or I could have proposed a definition of "preferable" as something that produces better agreements and convictions. Or perhaps I used "it's preferable" as a metaphor or starting point. I'll likely continue to use this metaphor.

None of the above compels me to accept anything morally; I am only bound by the non-contradiction of my own moral convictions—if that matters to me. It doesn't prevent us from seeing moral progress when looking at history on a broad scale, nor does it stop others from seeing progress when viewed from the future.

Some ethical treatises claim to prove objective moral truths. That is impossible. Others aim for shared moral propositions. That is possible. Some, both in the first and second groups, confuse the two. Others use metaphors that are not always clarified. There are attempts to outline the realm of moral propositions and distinguish them from other types of propositions, as well as efforts to trace the origins of our moral preferences and understand how malleable they are. These too would make for good ethical books, helping us understand one another. This

book, however, has more modest aims; it's not an ethical treatise in any of these senses.

With this book, I hope to accomplish three objectives. First, to develop the meaning of the phrases in this manifesto. Second, to offer a clearer approach to moral conflicts. And third, to show how certain tools from economic analysis, particularly the modern concept of utility, can support the first two objectives. This last point requires further explanation. All morality involves taking a stance in the face of moral conflict. For example, the principle "you shall not kill" has never meant "you shall never kill," but rather that exceptions should be kept to a minimum. Some people accept killing in self-defense, for revenge, or to prevent further deaths. All of these people can still accept the principle or commandment and even regard human life as sacred, but only in a metaphorical sense. Those who kill for pleasure or personal gain exclude themselves from this commandment. A similar statement can be made about other principles, such as "you shall not steal" or "you shall not lie." Economic analysis has developed tools to examine how scarce resources are allocated for competing purposes. Adhering to moral principles implies pursuing competing ends. The desire for multiple ends forces us to weigh them when they conflict. Different people will weigh principles differently. In general, the solution is not to impose one principle over others but to incorporate all, albeit none in absolute terms. In specific cases, however, one principle may prevail. Analysis and reflection help us to remain consistent in weighing these principles. There are many ways to be consistent.

Being reflective and consistent helps in developing one's moral preferences and applying them. It also aids in forming agreements and compromises among individuals who accept reflection and consistency. There are moral conflicts where compromise is neither possible nor desired. Internationally recognized human rights, for example, have largely become non-negotiable. In democratic societies governed by the rule of law, moral conflicts are almost always better resolved through

discussion and political compromise, rather than through force or narrow majority decisions. This book aims to encourage the presentation and discussion of such conflicts.

Chapter 1
Homo economicus, Homo ethicus

1.1 Classical utilitarianism

Jeremy Bentham (1748-1832) is regarded as the father of utilitarianism. His approach is perhaps best illustrated by one of the great moral issues of his time: the abolition of slavery. Although today the matter is settled, with the freedom of all people guaranteed by the Universal Declaration of Human Rights, slavery in Bentham's era was defended with elaborate arguments. Without delving into all of them, I'll present two key examples.

The first argument is religious. Many Christians found justification for slavery in the Bible, where there are numerous references to slave owners and servants, with no moral condemnation in the text. This arrangement is portrayed as normal in the eyes of Jehovah or the Lord—the most common names for the god of the Bible in the Old and New Testaments, respectively.

The second argument is racial and stems from observations of the varying levels of material development among societies around the world. This led many, particularly in more developed societies, to believe in the existence of races with different degrees of human characteristics. These individuals believed that some races possessed greater intelligence, reason, moral virtues, courage, and so on, while others were deemed inferior and could be tutored or even enslaved by the superior ones. John C. Calhoun, a politician and philosopher from South Carolina, even claimed that "never before has the Negro race of Central Africa, from the dawn of history to the present, achieved such a civilized and advanced condition, not only physically, but also morally and intellectually." [1]

Others pointed out that slaves in the southern states were treated better than poor immigrants in the northern states.

Both arguments rely on external criteria. If the criterion is accepted—whether the authority of a book in one case or racial superiority in the other—then slavery is justified. The counterarguments were also grounded in external criteria. The same Bible that provided arguments for slaveholders also offered support to abolitionists, particularly in the New Testament. A large compilation of texts, written by authors across different eras to support various interpretations of a religion, naturally allows for multiple readings. In the U.S., the Quakers, abolitionists since slavery was permitted in England, were especially vocal. Although religious arguments were the most common, other abolitionist arguments were also made, including appeals to the founding principles of the American nation and to the unnatural power imbalance slavery created between master and slave. Harriet Beecher Stowe's Uncle Tom's Cabin, published in 1852, provides a comprehensive overview of such arguments[2]. Despite its massive influence at the time, the novel has since been criticized for its portrayal of stereotypical characters, especially the figure of the "good Negro." A final set of arguments, more philosophical in nature, appealed to natural law, asserting that freedom is an inherent human right.

Bentham argues that moral judgments should not be based on external criteria, but directly on how they affect human welfare and happiness. Slavery is not wrong because of any particular interpretation of history, biology, natural law, or religion, but because it generates more suffering than benefit. According to Bentham's utilitarianism, we should tally the utils[3]—the measure of welfare or utility—gained by the slave owners and subtract the utils lost by the slaves. By observing the total, we would see the negative balance of slavery compared to its absence. This, Bentham claims, should be the foundation of morality: counting all happiness and all suffering to establish the fundamental axiom of utilitarianism, which he expresses as follows: "It is the greatest happiness of the greatest

number that is the measure of right and wrong,"[4] and "the obligation to minister to general happiness, was an obligation paramount to and inclusive of every other" (Bentham, 1776 [1988]).

Three principles underpin this early utilitarianism:

(i) All happiness and all suffering must be accounted for.
(ii) Everyone is equally capable of experiencing suffering and happiness.
(iii) Suffering and happiness can be measured and summed.

Principles (i) and (ii) imply that all human beings should be considered equally. Building on this, and adding principle (iii), Bentham concludes that all human beings, regardless of sex, race, social class, or any other condition, deserve equal rights. Furthermore, he advocates for divorce, promotes all forms of liberty—including economic freedom—and calls for the abolition of slavery, the death penalty, corporal punishment, child abuse even within the family, and the criminalization of homosexuality. He also proposed improving prison conditions and was even an early advocate of animal rights, famously writing: "The question is not, Can they reason? nor Can they talk? but, Can they suffer? Why should the law refuse its protection to any sensitive being?" (Bentham, 1789 [1948]).

No other philosopher of his time advanced as far as Bentham did, coming so close to the modern liberal view of what is morally unacceptable in society. For example, only a few years earlier, Immanuel Kant (1724–1804), with his categorical imperative, neither could nor would reach such conclusions. In fact, on many of these issues, he held opposing views, accepting unequal rights for women, serfs, and illegitimate children, the inferiority of certain races, and the immorality of homosexuality, to name a few. Yet, Kant is often celebrated as one of the greatest moral philosophers in history, while Bentham's utilitarianism is criticized as being insufficiently founded. One reason for this is that many philosophers tend to base morality on abstract principles without

considering the practical consequences of different moral positions and choices. Knowing the principles, they believe, allows us to solve moral dilemmas just as we solve problems of motion in physics by knowing the laws of mechanics. This is one of the most frequently admired aspects of Kant's work. Despite the poor outcomes presented by the Prussian philosopher[5], he is excused, much like Galileo is excused for not having developed the theories of Newton, Maxwell, or Einstein[6]. In the following chapters, particularly in section 2.6, I'll explain the impossibility of this Kantian approach and, consequently, the weakness of this argument against utilitarianism.

Below, I present further criticisms of Bentham's approach, along with an analysis of their validity. I anticipate that most of these criticisms are unfounded, rooted in a narrow view of utilitarianism or, worse, a misunderstanding of what constitutes a valid critique.

1.2 Criticism

Universality

Bentham is often criticized for abandoning the idea of a universal ethics. The most significant flaw attributed to his utilitarianism is not the difficulty of measuring happiness through the concept of utility, but rather the concern that the moral option yielding the greatest overall happiness may be poorly defined and subject to circumstances. For instance, in the case of slavery, what if the happiness of the masters in a society were substantial enough that, in a utilitarian calculation, it outweighed the suffering of the slaves? Should we then accept slavery in such a society? Doesn't this present a clear argument that utilitarianism cannot serve as a reliable moral guide?

The answer to the last two questions is negative. First, we could only say that this critique is valid if two conditions are met: (i) we already know from some other source that slavery is universally wrong, and (ii) there exists a

society where total happiness is greater with slavery than without it. However, we neither know of nor can conceive of any human society where these conditions are met. Perhaps in a society of intelligent ants or in the Borg collective from *Star Trek*[7] such a situation might arise, but not in a human one. Since condition (ii) is not met, it's unnecessary to refute the *a priori* assumption in (i), although we'll address it in the following chapters.

The fact that the anti-slavery argument could theoretically be rejected is not a weakness of utilitarianism, but rather one of its strengths: it can be empirically tested, and it has not been disproven. This is akin to Popper's concept of falsifiability in science, which, as we know, is one of the reasons science advances in knowledge while superstition doesn't. According to Galileo, all bodies fall with the same acceleration. It might turn out differently in some experiments, but it hasn't. Similarly, Bentham posits that in all human societies, happiness is greater in the absence of slavery. It might turn out differently in some societies, but we have not encountered such a case. When an experiment or observation either confirms or contradicts a theory, it points us in the right direction. Without empirical guidance, we don't know which way to go, and thought becomes stagnant. This happens in both science and ethics when we believe we have found unquestionable postulates.

Whether freedom is a good depends on whether we, as humans, deem it so. This is a simple idea, yet this subjectivity is rejected by those who insist on seeing freedom as an objective good, as they believe only then can freedom be properly justified. While it's their choice to theorize about morality in this way, it's neither a necessary choice nor has anyone succeeded in establishing such objectivity without resorting to *ad hoc* arguments, which ultimately require one to assume what one seeks to prove.

Five lives versus one life

Another common criticism of utilitarianism revolves around its supposed answer to the moral question of whether it's acceptable to sacrifice one

person to save five. The criticism assumes that, from a utilitarian perspective, the sacrifice is always justifiable—after all, isn't five lives more than one? Since this position is broadly viewed as morally repugnant, the criticism appears devastating. Furthermore, it promotes the idea that ethical principles must objectively affirm the sanctity of human life, so such dilemmas can be resolved swiftly without further debate[8].

However, this line of criticism is misleading. The claim that utilitarianism advocates for sacrificing one to save five assumes that the value of those lives is the only factor to be considered, which is not necessarily true. Consider a society where, every time five people need organ transplants and there are no available donors, a random individual is captured and sacrificed for their organs. Such a society would be unlivable. Suspiciously, the organs of politicians' or doctors' relatives would rarely appear on the operating table, and people would invest heavily in avoiding the patrols hunting for unwilling donors. These and other dysfunctions would arise, each with added costs that make the practice untenable.

This arbitrariness and the societal damage it causes allow us to reject these generalized sacrifice scenarios. Other contexts and assumptions, however—whether justified or flawed—could influence reactions to different versions of the dilemma. Many impose significant welfare costs that outweigh the benefit of saving five lives, though not all of them do.

Unlike the case of slavery, there are conceivable—and sometimes historically real—situations in which this moral dilemma has arisen, and the decision has been made to sacrifice a life. These situations typically occur in very limited, extreme circumstances, with no broader consequences for society. Consider a hypothetical example: a spacecraft is divided into two isolated modules. One module contains five astronauts, and the other contains only one. After an accident, if nothing is done, the five astronauts will die. The only way to save them is to

jettison the second module, condemning the lone astronaut to death. In such a scenario, it's no longer obvious that the ethical choice is to let the five die.

Similar circumstances can occur in shipwrecks, wars, mountain climbing, or rescue operations after natural disasters. Utilitarianism might not necessarily dictate that the right course of action is to sacrifice a life, as it acknowledges that there are more factors at play—such as the stability of social norms or the educational impact on children. In very specific cases, however, calculations may indeed favor sacrifice, but that's as far as utilitarianism goes.

An *a priori* position of never sacrificing a life is insensitive to context, including cases of self-defense, where killing an aggressor to save five family members may align with human moral intuition. Furthermore, this rigid position introduces a problematic distinction—often not clearly defined—between acting through action and through omission. For instance, is failing to follow a pre-established protocol an action or an omission?

Information

Another criticism of utilitarianism pertains to its consequentialist nature. Critics argue that since we cannot know all the consequences of our actions, utilitarianism lacks sufficient data to perform its calculations. However, this criticism is fundamentally flawed. Decisions are made based on the information available at the time. If anything, a lack of knowledge highlights the need to seek more information through data and theories about how the world operates, allowing for better predictions of the outcomes of our actions. In the absence of complete data, relying on principles as rules can be seen as employing a broad theory or hypothesis that serves as the best approximation of a given moral issue until more information becomes available. For proponents of *a priori* ethics, this may suffice. However, for utilitarianism and any other

ethical framework that values consequences, this represents only a provisional starting point that can be refined over time. Developing a robust moral theory requires extensive scientific research rather than mere armchair theorizing guided solely by intuition. Failing to evaluate consequences reduces *a priori* positions to simplistic shortcuts and reflects a mental laziness stemming from an unwillingness to seek the necessary data for evaluating complex situations.

Intentionality

Some critics argue that utilitarianism fails to account for the intentionality behind actions. For example, philosopher Bernard Williams contends that the death of a person by accident is considered as morally bad as murder, despite the latter being far more reprehensible. Like many criticisms of utilitarianism, this one stems from a limited understanding of the theory. While the loss of a life may eliminate the same amount of utility in society, it's not the sole consequence. A murderer who remains at large will create a loss of utility for citizens, who will experience increased fear. A society that punishes intentional homicides more severely than unintentional ones will likely produce greater social welfare than one that treats both equally. On one hand, such a distinction discourages the commission of murders; on the other, it avoids being overly punitive toward individuals who cause unintended accidents, as these individuals don't need to be deterred from actions they don't intend to commit. Additionally, there are various scenarios between accidental death and murder, such as reckless behavior or negligence. Utilitarianism can accommodate these nuances in intentionality when assessing the actions of those causing death.

Negative responsibility

In his essay *A Critique of Utilitarianism* (1973), Bernard Williams presents a pair of thought experiments designed to challenge utilitarianism and other consequentialist ethical theories. The first thought experiment involves a man who is a chemist, husband, and father in need of a job. A

friend, who works at a chemical company that develops products for the arms industry, offers him a position in his department. Although the man opposes the development of chemicals for warfare, his friend argues that if he declines the job, it will go to someone less scrupulous who will produce chemical compounds with a lower degree of control.

The second thought experiment is Williams' most famous. In this scenario, a captain in one country is about to execute twenty rebels. As a gesture to honor a foreign guest, the captain offers the guest the privilege of choosing one rebel to kill, allowing the others to go free. If the guest declines the invitation, the captain will proceed with the executions as planned. The rebels urge the guest to accept, but he is horrified at the thought of killing someone he bears no animosity toward, leaving him paralyzed by the weight of the decision before him.

For Williams, these cases of negative responsibility—where a person is held accountable for the consequences of not performing actions that others ultimately take—represent an insurmountable criticism of utilitarianism. He argues that the mental state of the chemist in the first scenario and the guest in the second must also be considered, not just the ensuing consequences. It is challenging to accept that the responses in both cases would be almost automatic, without at least questioning whether there is something less obvious in that acceptance. Can anyone be held morally responsible for actions committed by others in such situations? While it's true that the mental health of the protagonists is also a consequence of their decisions, this hardly compensates for the lives that will be lost. Williams doesn't provide a definitive answer, except to suggest that we may refrain from labeling a person's decision of not to take any action in these scenarios as immoral, even if that decision contradicts the utilitarian standard.

These thought experiments share some similarities with the cases discussed earlier, but they differ fundamentally. In the earlier cases, the deaths resulted from accidents, whereas in Williams' examples, they are

caused by third parties, whose moral responsibility should not be transferred to the decision-maker. The response to this criticism mirrors the arguments presented there, but an additional point can be made. Contrary to Williams' assertion, it's not always the case that utilitarian principles necessitate acceptance of actions in these examples. Beyond the consequences for human lives and the mental state of the decision-maker, there is another crucial factor: the validation of the emotional blackmail to which the individual is subjected. If accepting the proposal encourages further instances of blackmail in the future, while declining it discourages such behavior, then we have a situation where utilitarianism doesn't necessarily command accepting the emotional blackmail and aligns with Williams' position.

Equality

In classical utilitarianism, only the total utility matters. A quantity of 100 units of utility distributed equally among 20 people is considered the same as 100 units of utility concentrated in the hands of one person, leaving the other 19 with none. John Rawls, one of the most influential philosophers of politics and justice, was particularly critical of utilitarianism for this reason[9]. If one of the worst-off individuals in society loses one unit of utility so that two units can be given to one of the better-off individuals, the resulting distribution would be considered superior according to the utilitarian criterion.

This argument is similar to the one about slavery discussed in section 1.2, but stronger. To consider slavery as a critique of utilitarianism, one had to make the unlikely assumption that if a person transitions from being free to being enslaved by someone else, total utility might increase. Now, to use inequality as a critique, it suffices to conceive that society could change in such a way that a poor individual sees their utility decrease by less than the increase experienced by a wealthy individual. This issue is more complex than it may initially appear, because what we distribute are resources—not utility. For example, imagine we have an income of

100 to divide among four individuals. We could allocate 25 to each, yielding a utility of, say, 5 per person, for a total utility of 20. Now consider what it would take to generate the same total utility in a highly unequal distribution. Suppose that giving zero units of income results in such poor living conditions that it produces zero utility. In that case, we might give zero to each to three individuals (yielding zero utility) and allocate all the income to the fourth. To reach a utility of 20 in that one individual—assuming utility increases with income but at a diminishing rate—we might need to give him as much as 400 or, in any case, much more than the available 100. In this scenario, the unequal society would require a total income of 400 to match the total utility that an equal distribution achieves with just 100. Of course, the total income of 400, divided equally among the four individuals will yield a higher total utility. In other words, as long as utility increases less than proportionally with income, utilitarianism presents a strong case in favor of equalitarian societies.

Still, situations like the one outlined in the previous example can happen if society acquires more resources—through new discoveries or technological advancements—that are distributed toward the wealthy individual, while the poor individual not only misses out on those resources but also sees their chances of accessing previously available resources diminish, for example, because their job becomes obsolete. Although not as strong as it looked at first glance, this is a valid critique, as long as we can establish that utility equality is a desirable principle in cases where utilitarianism violates it like in this example. In section 6.4 I'll present a more exhaustive discussion on this.

However, this will not be a valid critique of modern utilitarianism, which is based solely on each individual's preferences, without external criteria. Modern utility theory doesn't sum utilities to make a social evaluation of one society over another. The most it can express in this case is what individuals themselves would say. That is, it would highlight that the poor individual would not want to transition to the more unequal society,

while the rich individual would. Any other person would also be free to hold their own moral preferences on the matter. From this clash of opinions and the actions society takes following its debate, there will be consequences for the example in question, which, once again, each person will evaluate in their own way. This might seem insufficient, but in the rest of this book, I'll aim to show that it's not and that significant progress can be made—even on issues of inequality—with a modern conception of utilitarianism. In particular, in sections 6.3 and 6.4, I'll demonstrate that it can reconcile the utilitarian perspective with Rawls' concerns about inequality.

The utils

The most compelling criticism of utilitarianism centers on the impossibility of observing "utils," the measure of utility. While this aspect of the theory cannot be defended, it can be modified to retain much of its core ideas and conclusions. This involves substituting a cardinal and objective measure of utility with an ordinal and subjective one, as modern economic theory does. This revised theory of utility only needs to postulate the existence of minimally consistent preferences, without reference to utils or other constructs. From this foundation, it demonstrates that each person can coherently rank different situations according to their own preferences, which is what we refer to as ordinal and subjective utility.

The theory of utility has significantly advanced since Bentham and the classical utilitarians two centuries ago, primarily because it has been articulated in precise terms. By exploring the implications of the theory and cross-checking it with real-world observations, we have been able to identify which parts of the theory are viable and which are flawed or poorly formulated, as well as how the theory can be improved. Specifically, while the measure of utility in utils is incorrect, it can be replaced with a preference-based utility function that doesn't rely on them.

This shift propels ethics grounded in utilitarian principles forward, instigating a kind of revolution: by substituting the nonexistent utils with actual individual preferences, the new utility framework compels us to transition from an objective interpretation to a subjective one. No other ethical theory can claim such significant advancements.

1.3 Utilitarianism based on modern utility theory

As the conception of utility evolves, so too does the idea of utilitarianism. However, in discussions of ethics, only the classical form of utilitarianism is typically analyzed and critiqued, perhaps because the modern version remains underdeveloped or has been relegated to the margins of contemporary philosophical discourse, primarily within economic analysis. Even among modern philosophers who identify as utilitarians, few explicitly adopt the modern utility function, even though they also don't adhere to the classical one. One notable exception to this trend is the debate between John Rawls and John Harsanyi, which we'll explore later in Chapter 6. However, this controversy doesn't appear to have significantly increased awareness of utilitarianism based on modern utility theory. In the following chapters, I'll develop these ideas further, to demonstrate what modern utilitarianism can contribute to the field of ethics.

Note that I have written "what modern utilitarianism can contribute." This indicates that this book is neither a treatise on ethics nor does it claim that utilitarianism will provide definitive answers to ethical dilemmas. No single theory offers those answers. At best, we can propose a system that presents various problems in an organized manner, making the coherences and contradictions of certain preferences, discourses, and actions clearer. This framework will facilitate individual decision-making and the negotiation of societal choices. Other ethical theories can also serve this purpose.

This new utilitarianism encompasses a set of ethical approaches characterized by the following elements: (i) reliance on individual moral preferences, (ii) the inherently subjective and often non-measurable nature of these preferences, measurable only in specific, axiomatically justified cases, (iii) the need to balance diverse moral goods and principles, (iv) the use of expected utility (or suitable alternatives) for analyzing situations of uncertainty, (v) a pragmatic approach to negotiations, (vi) the acceptance that one's own moral attachment to negotiation outcomes is not required, (vii) the application of reason to clarify preferences and to design processes for negotiation and preference aggregation, (viii) reliance on the best available scientific knowledge to predict the consequences of actions and agreements, and (ix) openness to revising preferences and negotiation approaches in light of new data. The proposed neo-utilitarianism is specifically defined by the combination of points (i) to (iv), although other ethical frameworks may share some of these and additional elements.

Throughout the book, I'll develop numerous examples of "ought to" analyses based on modern utility theory, as stated in points (i) through (iv), but I'll also incorporate the others. As a matter of fact, some of the examples don't make an explicit reference to modern utility theory. In sections 6.1 and 6.4 I'll revisit the case of slavery, presented in this introductory chapter, after providing a more detailed explanation of this theory.

I'll conclude this introduction with a simple example that can be understood in light of what has been discussed so far, which I believe will effectively illustrate both the differences from classical utilitarianism and its versatility in clarifying ideas that have often posed challenges for some modern ethical theories.

I have a friend in need, and I am considering giving him a lump sum of money. Before I do so, I realize that there are many people who are more in need than he is. Is it ethical to give the money to my friend instead of a

stranger? I believe that no inhabitant of planet Earth would even question this, and if someone did, they would likely answer in favor of helping the friend. However, ethical theories grounded in some form of objectivism or a neutral, universal perspective struggle to accept this. Classical utilitarianism would argue that it's preferable to help the stranger unless not assisting my friend would cause me significant distress that outweighs the difference in happiness between the stranger and my friend in the case of choosing one over the other.

Let's consider this further. If I increase the amount of money I am willing to give to my friend, reduce his urgency of need, and diminish the distance between myself and the stranger, we might eventually reach a scenario where I am wealthy, my friend is seeking money for a frivolous desire, and the stranger is a neighbor I see occasionally—someone who could solve a significant problem with a small amount of assistance. In this case, most people would likely conclude that the ethical choice would be to help the neighbor instead.

At some point along this continuum, I'll have changed my decision. No ethical system can definitively indicate when I should stop helping the friend and begin assisting the stranger. The best we can say is that different individuals will make this transition at different moments. The only clear instances we can objectively identify as immoral are the most extreme cases that contradict the above reasoning.

In modern utilitarianism, we begin with the moral preferences of each individual and then explore how we want to organize society and what types of actions we wish to encourage or discourage. A society that doesn't morally condemn helping a friend over assisting a stranger—except in extreme cases—will generally be preferred by everyone. However, at the same time, individuals will also desire that public aid be provided impartially. This natural solution is not only feasible within utilitarianism, but it also emerges logically from individual moral preferences.

This doesn't imply that utilitarianism validates individual moral preferences in any absolute sense. In particular, it doesn't mean that being a utilitarian requires me to accept the majority or prevailing moral views in my society. I maintain my moral positions and, if they contradict those of my peers, I'll attempt to persuade them or, if necessary, advocate and fight for my beliefs. What this example illustrates is that utilitarianism facilitates the clarification of one's moral consistencies and enables negotiation among individuals regarding the design and approval of various policies.

Modern utilitarianism allows for a clear distinction between what is morally acceptable to an individual, what is acceptable to others, and what is acceptable as public policy. When faced with the same problem, each perspective can provide a different answer without contradicting the others.

Notes and comments

[1] This quote and others like it can be read in his speeches, collected in Calhoun (1843).

[2] Stove (1852[2002]).

[3] Although the term *util* as the unit of measurement for utility came after Bentham's time, I'll use it here for the sake of simplicity in exposition.

[4] The phrase "the maximum happiness of the greatest number" doesn't accurately capture Bentham's utilitarian criterion. For instance, consider a society of 100 people with 70 % experiencing high happiness and the remaining 30 % enjoying medium-high happiness. Is this society better or worse than one where 90 % of the population experiences very high happiness and the remaining 10 % has medium-low happiness? This definition fails to differentiate between the two societies. If we could measure happiness in

utilitarian terms, we could draw clearer distinctions. For example, if very high happiness corresponds to 10 utils, medium-high to 8, and medium-low to 6, the first society would accumulate a total of 940 utils, while the second would have 960. However, if medium-high happiness were to yield 9 utils, the first society would be deemed superior. Therefore, this quotation from Bentham should be understood as a vague but illustrative summary of his theory, indicating that happiness increases along two dimensions: by enhancing the happiness of individuals and by increasing the number of individuals who experience happiness.

[5] After developing his theory of morality in *Groundwork of the Metaphysics of Morals* (Kant, 1785[2012]) and *Critique of Practical Reason* (Kant, 1788[2012]), Kant applies it in *The Metaphysics of Morals* (Kant, 1797[1996]). The second part of this book, dedicated to the *Doctrine of Right*, is arguably one of the most disappointing sections one may encounter in a treatise on ethics. After promising a moral framework that is apodictically deduced from the categorical imperative, readers are instead confronted with vague concepts, *ad hoc* principles, and arguments lacking rigor that ultimately reinforce Kant's intuitions, which, unsurprisingly, align too closely with the conservatism of the religious and political climate of his time. The following examples in the initial paragraphs of this section illustrate this point.

Vague concepts: “The united will of the people”, “Civil independence”.

Ad hoc principles: The attributes of citizens are “legal freedom”, “civil equality” and “civil independence”. That they are *ad hoc* principles says nothing about their desirability or goodness, but only that they are not derived by logical deduction from the foundations of Kant's own metaphysics of morals. Other *ad hoc* principles that permeate the whole book are appeals to what Kant claims is a supposed natural right.

Argumentation without rigor:

> The only qualification for being a citizen is being fit to vote. But being fit to vote presupposes the independence of someone who, as one of the people, wants to be not just a part of the commonwealth but also a

> member of it, that is, a part of the commonwealth acting from his own choice in community with others. This quality of being independent, however, requires a distinction between active and passive citizens, though the concept of a passive citizen seems to contradict the concept of a citizen as such. The following examples can serve to remove this difficulty: an apprentice in the service of a merchant or artisan; a domestic servant (as distinguished from a civil servant); a minor (naturaliter vel civiliter); all women and, in general, anyone whose preservation in existence (his being fed and protected) depends not on his management of his own business but on arrangements made by another (except the state). All these people lack civil personality and their existence is, as it were, only inherence (Kant, 1797).

Note the arbitrary definition of independence, which is presented as a binary concept—either yes or no—and categorizes self-employed individuals or public workers as independent, in contrast to employees. This definition is particularly arbitrary when distinguishing between a state employee and a private-sector employee. Following this logic, a doctor working in a public hospital is deemed independent, while a doctor in a private hospital is considered dependent. Moreover, a freelancer providing services to the same private hospital is classified as independent again.

I won't speculate on why Kant chose to use the examples of an apprentice and a servant rather than a skilled employee like a physician, whom he may have intended to classify as independent. Regardless, this approach results in an arbitrary distinction between different types of employees. Additionally, the identification of a dependent citizen—according to this criterion—with one who is forced to serve another is both arbitrary and clearly misleading. Can someone who chooses to serve another while maintaining autonomy truly be considered forced? Is there not a degree of choice involved?

Furthermore, consider the *ad hoc* principle that the ability to vote presupposes the independence described. Lastly, it's noteworthy that Kant categorizes all women as passive citizens without acknowledging the possibility of their independence in the sense he defines. This possibility was, in fact, a reality for

some women in various European societies of his time, such as certain widows who managed their deceased husbands' estates.

It can be rightly argued that Kant's racism, which pervades his writings on race, is even more disappointing because it contradicts his notion of ethical universalism. I have chosen not to include these aspects, as they fall outside the realm of his ethical theories and instead represent an ill-conceived attempt at biology and anthropology by someone lacking the necessary expertise. Nonetheless, the moral implications of these views are unmistakably clear.

[6] This analogy is, in fact, highly inappropriate. Galileo bases his conclusions on postulates supported by empirical evidence, whereas Kant asserts that his don't require such evidence. Galileo's conclusions are either clearly deduced from the postulates or they are not; he possesses the necessary elements to address certain problems in physics, or he doesn't. There is no room for debate here. In contrast, whether Kant's conclusions logically follow from his principles is a question still being debated more than two centuries later, particularly among those who believe such deductions are possible. Those who recognize this impossibility view Kant's arguments for what they are: appeals to a mixture of moral principles and analogies with a persuasive intent. Their focus is on reconciling modern moral positions with Kant's principles and analogies. Furthermore, Kant likens ethics to mathematics or logic, which begin with self-evident axioms to formulate and prove theorems that are formally deduced from them. It is important to note that Kant doesn't articulate such axioms for ethics (in fact, there is no self-evidence in the selection of axioms in logic or mathematics), nor does he deduce anything in a minimally formal manner.

[7] The *Star Trek* universe encompasses a well-known series of science fiction shows and movies. One of the civilizations depicted in this universe is the Borg, a collective that assimilates individuals from various species and interlinks them into a unified consciousness. In this society, individual autonomy and freedom are deemed expendable for the greater good of the collective. When an individual is separated from the collective, they often experience feelings of loneliness and disorientation, leading them to desperately seek reconnection. However, if they have not been assimilated for

long, they may have the opportunity to regain their individuality. In such cases, individuals describe their experiences with the collective positively, highlighting the sense of belonging to a shared intelligence, while viewing separation and the accompanying loneliness negatively—at least until they adapt to their individuality once more.

[8] In ethics textbooks, various iterations of the well-known *trolley problem* are commonly discussed. In these scenarios, a trolley is hurtling uncontrollably down a track, destined to run over five people unless it's diverted or stopped—options that invariably require sacrificing another individual. These dilemmas prompt reflection on the factors influencing our decision to redirect the trolley or not. The answers often hinge on the relationship of the sacrificed individual to the situation. For instance, it may be more acceptable to divert the trolley to a track where it will hit an employee of the company rather than a bystander. Other variations can significantly alter people's responses. For example, the dilemma might involve the option of stopping the trolley by throwing someone in front of it, which would certainly result in that person's death. The distinction between actively and passively causing a death often elicits different reactions, even among those who uphold the sanctity of human life in both scenarios. My aim is not to delve into these ideas—extensively covered in the existing literature—but rather to illustrate that many individuals may not always choose to sacrifice one person to save five. Additionally, numerous authors debate the ethical circumstances surrounding such sacrifices. Many discussions reveal that our moral preferences depend on factors beyond merely the lives at stake, focusing instead on identifying what that "something else" might be. Regardless of its nature, whether it enhances or detracts from societal utility, this consideration serves as a refutation of the critique of utilitarianism, which would otherwise reduce complex moral decisions to a simplistic acceptance of sacrifice.

[9] See, for example, Rawls (1988).