A Bestiary of Utility Monsters

Walter Barta

University of Houston

February 2021

Introduction

The concept of the Utility Monster offers an influential critique of Utilitarian theories, forcing us to consider different theoretical fixes to escape monstrous implications (Nozick, 1999, pp. 26-53; Kennard, 2015, p. 322). However, many different breeds, a whole bestiary, of Utility Monsters are identifiable, and each breed reveals something slightly different about what we find monstrous. When dissected in depth, we observe that some breeds are probably acceptable, whereas other breeds are indeed monstrous, though perhaps for slightly different reasons than Nozick thought. By breaking these taxonomies down, thus revealing strengths and weaknesses of these different breeds, we may see more clearly how the vicious versions can be dispensed with under reasonable assumptions.

Universal Utility Monsters

The original formulation of the Utility Monster was a critique of Total Utilitarianism, the view that we should maximize the Total Utility, the utility of the population, arguments summarized as the "greatest good for the greatest number" (Bentham, 2019, p. 7). Nozick's "utility monsters"[1] are thought experiments showing that Total Utilitarianism, Average Utilitarianism, and other theories[2] of utility are consistent with optimal outcomes in which a singular over-consuming person experiences most or all of the goods and the rest of the population experiences little or none (Nozick, 1999, p. 41).

The thought experiment shows us that, if Utilitarianism is true and Utility Monsters exist, then we must prioritize the monsters at the expense of everything else, including other Personal Utilities, the utilities of specific persons. Furthermore, even if Utility Monsters are not actual but merely possible, then we may be obliged to consider them anyway, perhaps even to create one to sacrifice ourselves to (Parfit, 1984, p. 388-89). What seems monstrous about these implications is that altruistic Utilitarianism seems to collapse into narcissism for the monster itself, and masochism for every other person (Frank, 2000). But, to add insult to injury, by definition, feeding the Utility Monster is obligatory; to not sacrifice oneself to the Utility Monster would be normatively monstrous. This is because, by definition, Utility Monsters have their Personal

[1] Nozick's original description: "Utilitarian theory is embarrassed by the possibility of utility monsters who get enormously greater sums of utility from any sacrifice of others than these others lose ... the theory seems to require that we all be sacrificed in the monster's maw, in order to increase total utility" (Nozick, 1999, p. 41).

[2] Nozick's other major target of criticism is John Rawls' maximin principle, "the greatest benefit of the least advantaged," which Nozick also believes is vulnerable to Utility Monsters, if the least advantaged behave as unconditionally demanding free-riders at the expense of every other contributing citizen (Rawls, 1971, p. 302).

Optimality, the optimal utility state of specific persons, coincide with the Total Optimality, the optimal utility state of the population. Thus, the Utility Monster shows that certain versions of Utilitarianism have counterintuitive consequences, and so Utilitarians should only accept versions that exclude Utility Monsters.

The original Utility Monster concept was Universal in the sense that it stipulated that "any" added goods for "all" persons give themselves to the monster (Nozick, 1999, p. 41). We can give this type of Utility Monster a more rigorous description here:

Definition: a person that has greater marginal utility than all other persons in all situations. (We have prefaced this "Universal" because it is such for all domains, which will contrast other forms.)

Mathematics: one arises when a person has a high-rate linear goods-utility function (high constant marginal utility). (We have also graphed the Utility Monsters in Figures 1-5, utility in a domain of commodities compared against a normal agent with diminishing marginal utility.)

Consequences: Other persons are never prioritized; the monster is prioritized indefinitely.

Real World Example: a human with a stoma implant in their stomach that can relieve food storage from the stomach as it accumulates is able to extract pleasure from food without their pleasure ever having to be bounded and diminished by feelings of satiety.

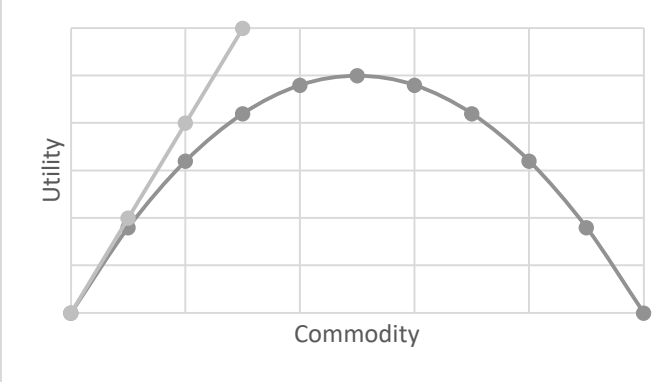

Figure 1: Universal Utility Monster

Good Features: the Universal Utility Monster does not seem to have any good features, besides for the good feature implicit to all Utility Monsters, which is that increasing the monster's personal utility maximally increases the total utility of the population.

Bad Features:

> **Sub-Zero Personal Utility**: none of the persons are ever entitled to achieve personal utility above zero, excepting the Utility Monster itself, which is itself entitled to the greatest possible utility within the domain of possibility, but less than its optimal (infinite) utility at infinity. The Total Optimal is never the personal optimal.
>
> **Indefinite Goods Requirement**: The marginal utility of goods persists indefinitely, such that an indefinitely increasing quantity of goods will always increase the total utility but never quite achieve Total Optimality.

Partial Utility Monsters

However, we can stipulate that there may be other kinds of Utility Monsters, non-universal (partial) Utility Monsters: kinds of utility monster that are monstrous only for some domains. These Partial Utility Monsters may themselves come in many different varieties depending upon where the threshold of monstrosity falls within the domain. Some thresholds may be preferable

to others. Notably, we will suggest that there are at least four different kinds of identifiable Utility Monster, which we will discuss in turn—Sub-optimal, Para-Optimal, Optimal, and Supra-Optimal. Dissecting out these variant breeds will help us diagnose what specifically we find so monstrous about Utility Monsters so that we can prescribe solutions.

Sub-Optimal:

Definition: a person that has greater marginal utility than all other persons in some personally sub-optimal situations.

Mathematics: one arises when a person has a low-rate linear goods-utility function (low constant marginal utility).

Consequences: In the presence of such a monster, other persons are briefly prioritized, until some diminished marginal utility, beneath their optimum, at which the monster's marginal utility exceeds their marginal utilities, after which the monster is prioritized indefinitely, for an infinite quantity of goods.

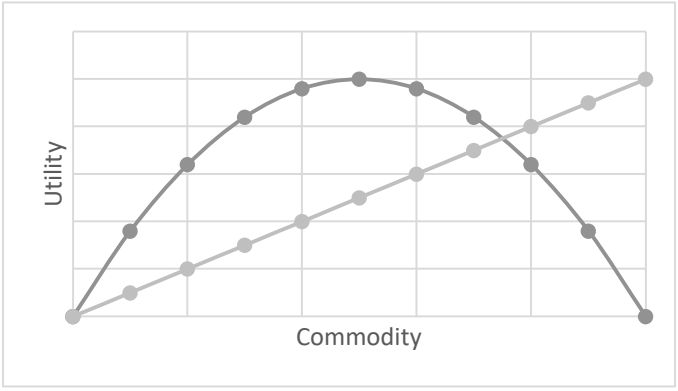

Figure 2: Sub-Optimal Utility Monster

Real World Example: a king permits his peasants to eat food but never enough to be entirely satisfied, and the king takes the rest of their food for his private court of nobles who gorge themselves ceaselessly to their indefinitely pleasure (Hobbes, Chapter XV).

Good Features:

> **Positive Personal Utility**: none of the other persons are ever condemned to subsist at zero or subzero utility. Rather, persons are entitled to some sub-optimal but positive level of personal utility. The Total Optimal is never personally miserable.

Bad Features:

> **Sub-optimal Personal Utility**: none of the persons are ever entitled to achieve their personal optimality, including for the monster itself, which is itself entitled to the greatest possible utility within the domain of possibility, but far less than its optimal (infinite) utility at infinity. The Total Optimal is never the personal optimal.
>
> **Indefinite Goods Requirement**: Same as above

Para-Optimal:

Definition: a person that has greater marginal utility than all other persons in some personally para-optimal (near optimal) situations.

Mathematics: one arises when a person has an asymptotic goods-utility function approaching some positive horizontal limit (decaying marginal utility).

Consequences: In the presence of such a monster, the monsters are prioritized until their marginal utility diminishes to equal some other persons' marginal utility, at which point those other persons are prioritized until their marginal utility diminishes to equal some other persons' marginal utility, indefinitely, always nearly but never quite optimal, at which point their marginal utilities are neck-and-neck with the monster's marginal utility indefinitely, without ever quite reaching their optimum, for an infinite quantity of goods.

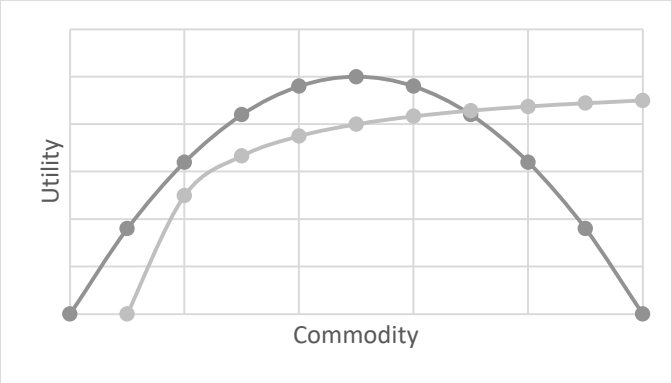

Figure 3: Para-Optimal Utility Monster

Real World Example: again, a king taxes his citizens progressively based on their satisfaction level, always taking more food the closer they get to total satisfaction, and takes the rest for his private court of nobles who do not get much extra pleasure from the extra food.

Good Features:

> **Near-Optimal Personal Utility**: none of the other persons are ever condemned to subsist at low levels of personal utility. Rather, persons are entitled to some nearly optimal level of personal utility. The Total Optimal is personally near-optimal.

Bad Features:

> **Sub-optimal Personal Utility**: Same as above. Though, this is better than for the Sub-optimal case because personal utility is near-optimal.
>
> **Indefinite Goods Requirement**: Same as above. Though, this is worse than for the Sub-optimal case because the indefinite goods required offer vanishingly small marginal utilities.

Optimal:

Definition: a person that has greater marginal utility than all other persons in some personally optimal situations.

Mathematics: one arises when a person has a negative polynomial goods-utility function (decreasing marginal utility); or a constant (horizontal) positive goods-utility function.

Consequences: In the presence of such a monster, the monster is prioritized until their marginal utility diminishes to equal some other persons' marginal utility, at which point those other persons are prioritized until their marginal utility diminishes to equal some other persons' marginal utility, until every person has reached their respective optimum at some finite quantity of goods.

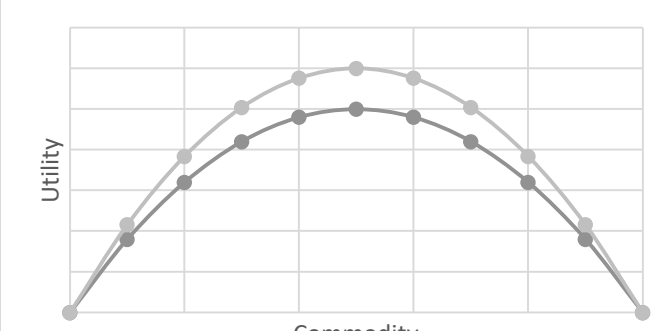

Figure 4: Optimal Utility Monster

Real World Example: a race of giant humans governs a race of miniature humans, and the giant humans extract the same amount of nutritive value from every pound of food as two miniature humans, although both races can reach satiety given enough food is available to them.

Good Features:

Optimal Personal Utility: persons are entitled to the optimal level of personal utility. The Total Optimal is personally optimal.

Finite Goods Requirement: The marginal utility of goods persists until some finite quantity, such that an indefinitely increasing quantity of goods will not indefinitely increase the total utility.

Bad Features:

Incidental Inequality: the monsters are prioritized over other persons under certain circumstances in which they stand to gain greater utility.

Supra-Optimal:

Definition: a person that has greater marginal utility than all other persons in some personally supra-optimal situations.

Mathematics: A Supra-optimal Utility Monster arises when a person has a positive polynomial goods-utility function (increasing marginal utility); or a constant (horizontal) positive goods-utility function.

Consequences: In the presence of such a monster, other persons are prioritized, until the diminished marginal utility at their optimums, after which point the monster is prioritized indefinitely.

Real World Example: a private club makes sure that every single existing member is completely satisfied, but then pulls a profit from its members indefinitely after. Furthermore, the club does not allow new members to join, and instead treats these new members as non-persons who are not entitled to anything.

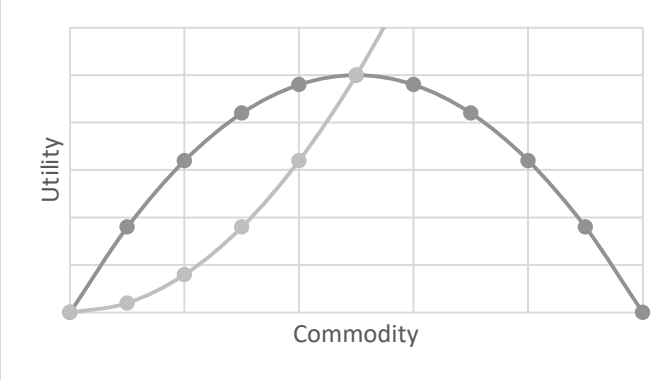

Figure 5: Supra-Optimal Utility Monster

Good Features:

Optimal Personal Utility: Same as above.

Bad Features:

Indefinite Goods Requirement: Same as above. Though this is better than the Para-Optimal case because the indefinite goods required offer increasing marginal utilities.

Zero Personal Utility (For New Persons): Once the marginal utility of the monster increases sufficiently, the marginal utility of the monster may exceed the marginal utility of any new persons added to the domain of consideration, such that the monster never permits the entry of new person into consideration.

Monsters and their Defeaters

In what follows, we will discuss the three major problems identified regarding Utility Monsters and will attempt to resolve each problem respectively. We will show that the first two seeming problems are intuitively acceptable, and we will show that the final problem can be excluded by properly specifying the normative domain.

Incidental Inequality

The Partial Optimal Utility Monster has the downside of obligating situations of Incidental Inequality. In any given instance of added goods, we might be obligated to give more of that good to Person A than to Person B until the marginal utility of Person A and B balances.

However, upon reflection, Incidental Inequality does not seem like a problem. Indeed, we routinely encounter situations of Incidental Inequality, such that they are trivial. On the banal end of situations, a slightly taller person might require slightly more food for sustenance than a slightly shorter person. On the emergency end of situations,[3] a dying person may require more urgent and stringent medical attention than a person suffering a minor injury (Singer, 1972). These temporary inequalities do not offend our intuitions; in fact, rejecting these cases on the grounds of unfairness would seem monstrous itself. Indeed, when reflecting upon the nature of action in the world, we can observe rather straightforwardly that differential action is required as a condition of doing anything, such that temporary inequalities of some duration and magnitude would have to be suffered even by the most egalitarian principles, if only just to avoid paralysis for long enough to get anything done (lest we find ourselves in the position of a Buridan's Ass-Altruist: someone unable to choose between which charity to give to who ends up forgoing the opportunity to give) (Aristotle, 2020, 295b). Generalized equality does not imply instantaneous equality at every point in spacetime. So, not only are personal Incidental sacrifices understandable, but they are also necessary in the limit case if only to avoid Buridan's Ass styled dilemmas.

Sub-Optimal Personal Utility:

The Partial Sub-optimal and Para-Optimal Utility Monster have the downside of obligating situations of Sub-optimal Personal Utility. For any given case of Total Optimal Utility, any given person might not be permitted optimal personal utility. For instance, at a given Total Optimum, Person A might be obligated to sustain a sub-optimal personal utility.

However, upon reflection, although Sub-Optimal Personal Utility seems like a personal problem, as a consideration of the Total Utility it does not seem like a problem. Indeed, we can easily imagine that some Total Optimums may permit and even require Personal Suboptimums. On the banal end of such situations, given finite amounts of food, a person might be obliged to not eat as much as they want so that they can share with another person. On the emergency end of such situations, given certain dangers, some number of persons might have to sacrifice wellbeing to protect the optimums of the rest of the population. These personal sacrifices to the total wellbeing need not offend our intuitions; in fact, while these circumstances may be lamentable,

[3] Some such emergency cases seem so egregious that it would seem monstrous for Nozick to object to them on the grounds of libertarian constraint (Nozick, 1999, p. 45).

they may not be preventable given material constraints in the world (Sidgwick, 1962, p. 162). Indeed, reflecting upon the nature of material tradeoffs,[4] we find no guarantee that all optimums will coincide for every person in every situation (Pareto, 1971). The default assumption should be that optimums can sometimes conflict (Miettinen, 1999, pg. 5). (Benefiting you may necessarily cost me, and vice versa.) Total Optimality does not imply personal optimality for every person in the population. So, not only are personal suboptimums understandable, they are also necessary in the limit case as a material constraint on material beings.

Indefinite Goods Requirements

Every Utility Monster (except the Partial Optimal) has the downside of requiring indefinite usage of goods. The problem more rigorously defined is that for any given case of monotonically increasing utility in a domain of increasing quantities of goods, there must always be some utility higher than the previous utility, which implies that there must always be some greater set of goods to pursue beyond the previous sets of goods. For instance, for a given population of indefinitely required goods, Person A and B should not be satisfied with x goods because they can be satisfied with x+1, and should not be satisfied with x+1 because they can be satisfied with x+2, etc.

Like the previous two problems, Indefinite Goods Requirements do not seem obviously bad. If one of something is good, ceteris paribus, isn't two of something better? In finite ranges, this is reasonable. However, unlike the two previous problems, extrapolated to the extreme, the requirement becomes a problem, as the logical conclusion of an indefinite transitive sequence of values is that the highest value must be infinite. For instance, Person A and B should not be satisfied with x+2 because they can be satisfied with ∞. Indefinite Goods Requirements end up implying Infinite Goods Requirements. This is a problem because acquiring infinite goods is a material impossibility.

However, this argument ad infinitum reveals the weakness of Indefinite Goods Requirements, that we might defeat it once and for all. Namely, in any given circumstance for any given set of persons, infinite goods requirements should be excluded from normative considerations as contradictions: an impossible necessity. One can see this as a special instance of "ought implies can" (Kant, 1793, p. 94). Because we cannot consume infinite material goods, we ought not consume infinite material goods, and so we ought to reject any normative systems that end up implying that we are required to do so, as do versions of Utilitarianism that permit of Utility Monsters. Accepting this principle eliminates all Utility Monsters (excepting the Partial Optimal ones) from normative consideration and thus all theories friendly to them, including unbounded Utilitarianism. Thus, contra-Nozick, we propose not a "moral side constraint", Kantian "inviolability",[5] but an alethic side constraint, Kantian impossibility, as a defeater of the Utility

[4] Nozick would surely be in agreement here, as the background of material tradeoffs is implied by his analysis (Nozick, 1999, p. 33).

[5] On Nozick's account, the problem arises when one is normatively required to "sacrifice" some persons to other persons, and his solution is to categorically exclude such occasions; thus, he departs greatly (which is by-and-large the content of *Anarchy, State, and Utopia*) with our interpretation of the problem and our proposal for a solution (Nozick, 1999, p. 45).

Monster (Nozick, 1999, p. 32-33). However, it is conceivable that any alethic constraint can be set aside if we give the utility monster the imperative to consume all goods *possible*, such that we stipulate its upper bound, and it fills the universe with its own utility and but then stops short.

If the alethic constraint is inadequate, there are two other reasonable considerations to make to exclude such monsters.

First, we might observe that such Utility Monsters are necessarily psychologically exotic. As a comparison class we can consider a normal psychological agent as one with a diminishing marginal utility function (Mises, 1998). Only the Optimal Utility Monster has a (normal) diminishing marginal utility. In contrast, a Universal or Sub-Optimal Utility Monster will have a constant marginal utility, implying a mind that oddly has no preference ordering and will approach an infinite utility. A Supra-Optimal Utility Monster will have accelerating marginal utility, implying a mind that oddly eventually approaches an infinite marginal utility. A Para-Optimal Utility Monster will have an always positive marginal utility, implying a mind that oddly has no negative marginal utility. If we can exclude some or all of these exotic psychologies then some or all Utility Monsters can be ruled out. It seems like Universal/Sub-Optimal monster minds require static psychologies, unaffected by any differential consumption, which seems impossible in a dynamic world; as well as infinite positive utility values, which seem impossible in a finite mind. It seems like a Supra-Optimal monster minds require psychological preference sets that include infinite positive marginal utility values, which seem impossible in a finite mind. It seems like Para-Optimal monster minds require psychological preference sets that are entirely positive, without any aversions whatsoever, which seems perhaps possible but wildly maladaptive since it seems unlikely that psychologies without aversions would ever have survived environmental dangers and evolved amidst such Darwinian pressures. Thus, by process of elimination, only acceptable kinds of Utility Monsters, the ones with Optimal (and perhaps non-Darwinian Para-Optimal) monster minds are possible.

Second, if excluding exotic psychologies will not do, we might imagine some non-Nozickian "moral side constraint". The solution, per Nozick, is to strongly stipulate that the individual is *always inviolable*, but we have shown this solution to itself be monstrous (see the "sub-optimal personal utility" section above); thus, contra-Nozick, we may propose instead to weakly stipulate that the individual is *not always violable*, a subtle but important distinction. Relative to others, it is not unconditional inviolability but conditional violability; it is not that we *cannot ever* use persons as means to our ends, but that we *cannot always* use persons as means to our ends. The universal utility monster feels so monstrous because it requires unconditional violations in an domain of indefinite goods requirements, which becomes acceptable when the goods requirements are made definite and the violations conditional.

Monsters' Treasures

A first upshot: by excluding infinite values, we have constrained the possible utility functions (the mathematics by which we assign values to persons and goods) consistent with our non-monstrous intuitions, narrowed from Bentham's unbounded types down to a narrow acceptable range: the utility function must not be monotonically increasing (Broome, 2004). Although this result may seem counterintuitive, it is consistent with many of the intuitions that we already have: to say that values are not monotonically increasing can be construed as a more technically

rigorous statement of Aristotle's ethic of moderation against greed (*pleonexia*): virtuous means and vicious extremes (Aristotle, 2011).

A second upshot: this excludes a wide range of utility functions as unacceptable. First, as a matter of public policy, we should reject as monstrous any normative standard that assumes or implies infinite values (for example Parfit's "Repugnant Conclusion") (Parfit, 2004). This might include investors demanding infinite growth models of the economy; this might also include free-riders demanding bottomless safety-nets. Second, given that future generations may design artificial agents and specify the utility functions governing those agents, we should decline to value any agents demanding monotonic increasing utility functions (Fisher, 2020). These might include institutions, like companies with high profit margins; these might also include machines, like algorithms bent upon perpetual optimization.

Conclusion

So, we have described how Total Utilitarianism is vulnerable to a bestiary of two different genera of Utility Monster (Universal and Partial), and that Partial Utility Monsters come in several different species, including: Sub-optimal, Para-Optimal, Optimal, and Supra-Optimal. We have shown that these breeds of Utility Monster impose three major problems: Incidental Inequality, Sub-Optimal Personal Utility, and Indefinite Goods Requirements. However, all three of these problems are defeated by some reasonable assumptions—we ignore these assumptions at our peril, as that way there be monsters.

Appendix A: Formulas

The following formulas describe Utility Monsters at utility (u), in terms of goods (x), for persons (i), given the existence one utility monster ($i = m$), and some set of personal optimums ($x_{i,opt}$).

A.1: Universal

$$\forall x \forall i \left(\frac{du_m}{dx}(x) > \frac{du_i}{dx}(x) \right)$$

A.2.a: Sub-optimal

$$\forall x \forall i \left(\frac{du_m}{dx}(x_{i,opt} > x) > \frac{du_i}{dx}(x_{i,opt} > x) \right)$$

A.2.b: Para-Optimal

$$\forall x \forall i \left(\frac{du_m}{dx}(x_{i,opt} \approx x) > \frac{du_i}{dx}(x_{i,opt} \approx x) \right)$$

A.2.c: Optimal

$$\forall x \forall i \left(\frac{du_m}{dx}(x_{i,opt} = x) > \frac{du_i}{dx}(x_{i,opt} = x) \right)$$

A.2.s: Supra-Optimal

$$\forall x \forall i \left(\frac{du_m}{dx}(x_{i,opt} < x) > \frac{du_i}{dx}(x_{i,opt} < x) \right)$$

Works Cited

Aristotle. *De Caelo*. Translated by C. D. C. Reeve, Hackett, 2020, Indianapolis.

Aristotle. *Nicomachean Ethics.* Translated by Bartlett, Robert C. and Susan Collins, Chicago: University of Chicago Press, 2011, Chicago.

Bentham, Jeremy. *An Introduction to the Principles of Morals and Legislation*. Anodos Books, 2019.

Broome, John. *Weighing Lives*, Oxford University Press, 2004, Oxford.

Fisher, Richard. "The intelligent monster that you should let eat you." BBC News, 13 November 2020*, https://www.bbc.com/future/article/20201111-philosophy-of-utility-monsters-and-artificial-intelligence.* Retrieved 12 February 2021.

Frank, Robert H. (June 2000). "Why Is Cost-Benefit Analysis so Controversial?" *The Journal of Legal Studies*, Volume 29, Number S2, June 2000, pp. 913-930.

Hobbes, Thomas. *Leviathan.* Edited by Rogers, G. A. J., Schuhmann, Karl. London: Bloomsbury Publishing, 2006, London.

Kant, Immanuel. *Religion within the Boundaries of Mere Reason*. Edited and Translated by Wood, Allen and George Di Giovanni, Cambridge University Press, 1998, Cambridge.

Kennard, Frederick. *Thought Experiments: Popular Thought Experiments in Philosophy, Physics, Ethics, Computer Science & Mathematics* (1st ed.). AMF, 2015.

Miettinen, Kaisa. *Nonlinear Multiobjective Optimization*. Springer, 1999.

Mises, Ludwig Von. *Human Action: A Treatise on Economics*. Auburn, Alabama, The Ludwig von Mises Institute, 1998.

Nozick, Robert. *Anarchy, State, and Utopia*. Blackwell Publishers Ltd., 1999, Oxford.

Pareto, Vilfredo. *Manual of Political Economy*. Translated by Ann Schwier, Augustus M. Kelley Publishers, 1971, New York.

Parfit, Derek. *Reasons and Persons*. Clarendon Press, 1984, Oxford.

Parfit, Derek. "Overpopulation and the Quality of Life". *The Repugnant Conclusion: Essays on Population Ethics.* Edited by Tännsjö, Torbjörn and Jesper Ryberg. Library of Ethics and Applied Philosophy: Vol. 15, Springer Netherlands, 2000, Dordrecht, pp. 7–22.

Rawls, John. *A Theory of Justice*. Belknap Press, 1971, Boston.

Sidgwick, Henry. *Methods of Ethics* (7th Edition). Palgrave Macmillan, 1962, London.

Singer, Peter. "Famine, affluence, and morality." *Philosophy and Public Affairs,* vol. 1, num. 3, 1972, pp. 229-243.