1 Introducing Utilitarianism

Utilitarianism is a historical tradition in moral and political thought. Although utilitarian themes are present in all philosophical schools – throughout the Western tradition since the Ancient Greeks, and also in early Chinese and Indian thought – modern utilitarianism is especially associated with three thinkers active in Britain between the late eighteenth and late nineteenth centuries: Jeremy Bentham (1748–1832), John Stuart Mill (1806–73), and Henry Sidgwick (1838–1900). Utilitarianism was a dominant mode of ethical thinking in Western philosophy in the early twentieth century. Although less dominant today, it remains very influential.

This Element is neither a historical account of the utilitarian tradition nor a standard textbook introduction to contemporary utilitarianism. Several other books already fill those niches admirably.[1] Instead, this Element explores the future of utilitarianism, asking where utilitarians' perennial preoccupations might lead in various possible futures. Section 1 introduces the utilitarian tradition and the approach taken in this Element. Section 2 argues that, in our present circumstances, the future should dominate our ethical thinking and that any adequate utilitarian future ethic will be collective and pessimistic. Section 3 outlines contemporary debates about the content and scope of well-being and asks how those debates might be transformed across a range of different possible futures. Section 4 addresses a number of puzzles in contemporary future ethics – especially Parfit's Repugnant Conclusion and Non-Identity Problem, asymmetries in procreative ethics, the destabilising impact of empirical and normative uncertainty, and existential threats of human extinction.

Any account of a tradition as rich and varied as utilitarianism is bound to be controversial. My aim here is not to defend any detailed exegesis (either historical or contemporary) but rather to draw out some central utilitarian themes. The defining feature of utilitarianism is that it bases its moral evaluations on *impartial promotion of well-being*. (As we'll see, different utilitarians evaluate different things: acts, rules, moral codes, social institutions.) Impartiality, promotion, well-being: these three key terms need unpacking. I explore promotion in section 2.2 and well-being in section 3. I begin with impartiality and its implications.

Utilitarians are committed to *impartiality.* In the famous phrase attributed to Jeremy Bentham: 'Everyone to count for one, and nobody for more than one.'[2]

[1] Good introductory textbooks on utilitarianism include Bykvist, 2009; de Lazari-Radek and Singer, 2017; Mulgan, 2007; Shaw, 1999. An excellent contemporary overview is Eggleston and Miller, 2014. An excellent recent historical overview is Schultz, 2017.

[2] The attribution goes back to Mill, 1963, vol. 10, p. 257. While it is often attributed to Bentham, this precise phrase is apparently not found in any of his extant writings. Perhaps the closest

Human well-being is equally valuable no matter whose it is. Following Bentham, utilitarians emphasise impartiality as a counterweight to the perennial threat of *egoism*. This threat is both practical and theoretical. We must guard against our natural tendency to give undue weight to our own interests, values, traditions, or perspectives or to *believe* what suits our interests, aligns our duties with our inclinations, confirms our prejudices, or otherwise enables us to think well of ourselves. As a result, utilitarians are especially suspicious of moral principles that allow us to privilege our own interests.

Many moral theories agree that we should treat persons impartially. (Indeed, many philosophers have built impartiality into the very definition of the 'moral point of view'. See, e.g., Baier, 1958; Hare, 1982.) But utilitarians go further in two ways. First, utilitarians are impartial between species – or, more generally, between *kinds* of beings for whom things can go well. Well-being is defined without reference to any particular species – in particular, without special reference to *Homo sapiens*. It is then an empirical question whether or not non-human animals matter. For instance, if – as hedonists argue – well-being is pleasure and the absence of pain, then all sentient animals matter, and they matter in exactly the same way as human beings.[3] This feature of utilitarian impartiality has notoriously radical implications for our treatment of animals. (It is not a coincidence that many leading figures in the animal liberation movement are utilitarians. See especially Singer, 1975.) I argue below that future utilitarians may face analogous challenges relating to the well-being of extra-terrestrial organisms or digital beings. And, as we shall see throughout this Element, the underlying commitment to impartiality has many other implications for utilitarian future ethics.

Modern philosophical utilitarianism comes in a bewildering variety of forms. We can illustrate these by starting with the following basic formulation.

Hedonist Act Utilitarianism (HAU): The right act in any situation is the act that maximises expected total net pleasure (i.e., the balance of pleasure over pain).

Although it appears simple, we can break HAU down into ten component claims:

approximation is the following passage from Bentham's *Rationale of Judicial Evidence*: 'every individual in the country tells for one; no individual for more than one.' (Bentham, 1838, vol. 7, p. 334.)

[3] Hedonists could still regard human well-being as more significant, but only insofar as humans are capable of greater heights of enjoyment or greater depths of suffering than animals. A human's pleasure in listening to a symphony or terror in the face of imminent torture might simply be greater than anything a spider can experience. But if the spider could appreciate Mozart, its pleasure would count as much as ours! We return to hedonism in section 3.1 and to animal welfare in section 3.2.

1 **Consequentialism:** Utilitarians insist that well-being is valuable. But there are many different ways to respond to the belief that something is valuable. Should we *honour* what is valuable? (Treating its instances with respect, reverence, or worship; seeking to protect or preserve them.) Or should we try to *embody* or *instantiate* particular values? (Should utilitarians recognise the value of happiness by seeking to be happy?) Consequentialists argue that we should *promote* value by aiming to bring about valuable outcomes. Contemporary moral philosophy treats utilitarianism as a species of consequentialism. In this Element, I will largely follow this assumption. However, I often question specific consequentialist claims, and in section 4.3 I explore explicitly non-consequentialist forms of utilitarianism.
2 **Welfarism:** Outcome value is determined exclusively by the welfare of individuals. (Alternatives include ecological values and perfectionist values.)
3 **Hedonism:** Individual welfare is determined exclusively by pleasure and pain. (Alternatives include preference satisfaction and objective goods.)

Consequentialism tells us to promote value, welfarism identifies value with optimal well-being, and hedonism delivers a metric for *the value of individual lives*. HAU then aggregates individual value along three dimensions: persons, times, and prospects (cf. Broome, 2004).

4 **Totalism:** Outcome value is determined by total welfare. (Alternative measures include average welfare and the distribution of welfare.)
5 **Temporal Neutrality:** The contribution that an individual life makes to the value of an outcome does *not* depend on when that life is lived. (Alternatives include temporal discounting and other temporally asymmetrical views.)

Aggregating across persons and times gives us a measure of *outcome value*. To measure the value of *acts*, HAU then aggregates over prospects.

6 **Expectation:** The value of an act is the sum of the value of each prospective outcome multiplied by the probability of that outcome occurring if the act is performed. (Alternatives include risk-aversion or risk-seeking.)

Having ranked acts according to their value, HAU then tells us how we should *respond* to act values:

7 **Maximisation:** The right act is the one with the greatest (expected) value. (The main alternative is *satisficing*, where agents are permitted to choose any act whose (expected) value is 'good enough'.)

Finally, HAU's initial decision to focus on *acts* itself combines three controversial claims:

8 **Act Focus:** The primary focus of consequentialist evaluation is acts. (Alternative foci include rules, motives, codes, outlooks, dispositions, institutions, constitutions, beliefs, etc.)
9 **Direct Evaluation:** We evaluate each act directly in terms of its consequences. (The alternative is to evaluate one unit *indirectly* in terms of a second unit chosen because of *its* (direct) consequences. For instance, rule consequentialism first directly evaluates codes of rules and then uses the optimal code to indirectly evaluate acts.)
10 **Individual:** The primary focus of consequentialist evaluation is the particular acts of an individual agent. (The main alternative is the collective evaluation of sets of acts performed by groups of agents.)

Each of these ten claims is a site of ongoing controversy within contemporary moral philosophy. Some claims are regarded as *essential* to utilitarianism. For instance, utilitarianism is often *defined* as the set of possible theories that combine consequentialism and welfarism. And, as I said at the outset, temporal neutrality is often regarded as a central utilitarian commitment. Rejecting any of these three components would amount to a rejection of utilitarianism per se. But that still leaves seven components: hedonism, totalism, expectation, maximisation, act focus, direct evaluation, and individual evaluation. Rejecting any of these claims counts as a move *within* utilitarianism.

I set maximisation and expectation aside for now. (I return to them briefly in sections 4.5 and 4.6.) I will focus on three remaining contested aspects of the utilitarian response to value.

1 **Individual welfare:** Is hedonism the correct account of well-being?
2 **Aggregation:** Is totalism the best way to go from the value of individual lives to the value of outcomes?
3 **Promotion:** Once we have ordered possible outcomes by their value, how should we respond? What is our focus of evaluation (acts, motives, rules)? And should we evaluate that item directly or indirectly, individually or collectively?

I address well-being and aggregation in sections 3 and 4.1 respectively. However, for reasons that will emerge as we proceed, I begin with the third set of questions.

2 A New Utilitarianism: Future-Oriented, Collective, Pessimistic

2.1 Why the Future Should Dominate Utilitarian Thinking

Utilitarians have long argued that our obligations to distant strangers are more onerous than most of us assume. We cannot discount well-being just because it

happens a long way away. Our obligations to distant future people are, if anything, ever more pressing than what we owe to present distant strangers. Our relations with distant future people are integral to our own moral community – and arguably to our present projects and achievements. Also, our impact on future people *includes the content of their moral outlook*. These connections will be a central theme of this Element.

Up until the late twentieth century, like most other moral philosophers, utilitarians focused primarily on obligations to contemporaries and/or people in the near future. This can seem surprising, because utilitarians have always recognised that the well-being of even the most distant future people is as important as our own. The general utilitarian commitment to impartiality extends to *temporal impartiality*. Human well-being is equally valuable no matter whose it is – *or when they live*. For utilitarians, the interests of future people count as much as those of present people. In other words, utilitarians reject *pure temporal discounting*.

The practice of discounting future harms and benefits is relatively uncontroversial as a proxy for uncertainty – or to accommodate the remote possibility that there will be no future people. (Humanity might be wiped out by an unpreventable asteroid strike, for instance.) There are also sound utilitarian reasons to discount the future *if* you are confident that future people will be richer than present people or that technological advances will leave them better able to exploit any valuable resource. (Of course, as we shall see, this argument is reversed if we expect future people to be worse off.) The controversial philosophical question is whether we should apply pure temporal discounting, where future happiness counts for less simply *because* it lies in the future. One common justification is that this pure time preference mirrors actual behaviour. We do discount future benefits both to ourselves and to others.

Discount rates have a huge practical impact. Climate change provides a striking illustration. One prominent sceptical argument holds that the *future* benefit of preventing climate change is not worth the *present* cost. We could do more good by devoting our present resources to the alleviation of present poverty. Even a modest discount of 5 per cent per annum makes it 'uneconomic' to spend even one dollar today to avert a global catastrophe in five hundred years' time. (To be worth spending a dollar today, the catastrophe has to cost $137,466,652,006 at that future date.) Different economists reach radically different conclusions on the basis of their divergent discount rates. (Compare, e.g., Stern, 2006 and Nordhaus, 2007, pp. 143–61.)

While pure time preferences are controversial among economists, most utilitarian *moral philosophers* reject them and embrace temporal impartiality (Cowen and Parfit, 1992). Unlike some non-utilitarians (e.g., Heyd, 1992),

utilitarians cannot simply set future ethics aside. So why did utilitarians ignore the distant future despite recognising that distant future people are just as important?

Earlier utilitarians sidelined the future for three main reasons: *optimism*, *similarity*, and *convergence*. Like other philosophers, they assumed (1) that future people will be better off than present people, (2) that the future will resemble the present in most morally relevant ways, and (3) that the interests of present and future people largely converge. Environmental crises and other recent developments undermine all three presuppositions. We must now recognise that future people might be worse off than ourselves, that they might inhabit very unfamiliar futures, and that their interests might conflict with our own. Utilitarians must now pay special attention to future ethics.

Traditional moral and political philosophy often presumes an *affluent future* that resembles the past in most morally significant ways, with the exception that future people will be better off than us. In particular, it is assumed that future societies enjoy what John Rawls calls 'favourable conditions' (Rawls, 1971, p. 178), where it is possible to establish liberal democratic institutions that meet all basic needs without sacrificing any basic liberties. And Rawls argues that if such institutions *are* established, then we can expect to see modest increases in living standards across generations. What is best for us is thus also good for our descendants.

An affluent future is possible. But it is not inevitable. In my book *Ethics for a Broken World*, I imagine a future broken by climate change, where a chaotic climate makes life precarious, each generation is worse off than the last, it is no longer possible to meet everyone's basic needs, and our affluent way of life is no longer an option (Mulgan, 2011, 2014a, 2015b, 2015c, 2015d, 2016, 2017, 2018a, 2018b, 2018c).

This broken future is *credible*. No one can reasonably be confident that it won't happen. It involves no outlandish claims, scientific impossibilities, or implausible expectations about human behaviour. Climate change – or some other disaster – might produce a broken future. I argue in sections 3.4 and 3.5 that even some of the brightest-looking futures have a potentially (very) broken flipside.

The credibility of the broken future undermines our three presumptions of optimism, similarity, and convergence. In a broken future, people are (by definition) worse off than ourselves. As I argue below, their moral challenges differ from our own. Finally, both the likelihood and the severity of the broken future may depend on *present* choices. In particular, we might be able to mitigate future brokenness by making present sacrifices. If so, present and future interests conflict – and one key task for future ethics is to balance these competing interests.

Of course, while it is credible, the broken future is not inevitable. The future might be much better or much worse. Affluent liberal society promises to meet all basic needs but not to satisfy all desires. Once basic needs and basic liberties are guaranteed, the primary focus of our affluent theories of justice is the distribution of resources that remain scarce *relative to desires*. In the broken world, the affluent promise is broken: not all basic needs can be met. At the other extreme lies the promise of a *post-scarcity* future where resources are not scarce even relative to anyone's desires (Mulgan, 2017). Like the broken future, this other extreme possibility cannot be ruled out.

It is tempting to assume that post-scarcity futures are ethically uninteresting. However, this is too quick an assumption. In a post-scarcity world, it is *possible* to simultaneously satisfy all desires. It does not follow that this happens automatically or permanently. Post-scarcity conflict is still possible for several reasons. First, powerful individuals whose desires are all *already* satisfied might oppose a new system designed to satisfy everyone else's desires as well. Why take the risk? (Post-scarcity technology, if concentrated in a few hands, might enable the few to easily dominate the many.) Second, people may reasonably disagree about how post-scarcity life should be organised. And we cannot dismiss such disagreements as trivial matters of taste. On any account of human well-being other than crude actual-present-desire-maximisation, some possible post-scarcity scenarios are (very much) better than others. It is good (other things being equal) if all desires are satisfied. But it also matters what those satisfied desires are *for*. Imagine a post-scarcity paradise where nanotechnology has produced 'cornucopia machines' capable of re-assembling air molecules to create any desired object (Stross, 2005). We can still ask what people will *do* with their cornucopia machines. Will they all descend into a drug-induced stupor or retreat into mindless virtual realities? (What do present people do with the potentially infinite resources of the World Wide Web?) Will anyone have the incentive or the drive to invent or explore *new* patterns to programme into cornucopia machines or new ways to spend their (now effectively limitless) leisure time? Cautionary tales of wishes granted by duplicitous literal-minded genies, Asimov-literal robots, and other post-scarcity fictions teach us that a world where everyone gets what they want could be shallow, unstable, or otherwise very grim. (In Stross, 2005, for instance, the arrival of cornucopia machines escalates existing social tensions into an all-out civil war.) Finally, unless we imagine beings with radically non-human desires, our post-scarcity world cannot literally involve everyone getting everything they want. People's strongest desires often ineliminably involve other people. And those desires inevitably conflict. (Consider Hobbesian desires for power or pre-eminence, the never-ending consumerist urge to keep up with the Joneses, or the desire for

a reciprocal and exclusive romantic relationship.) Once again, it matters how these post-scarcity conflicts are resolved.

We simply have no idea what the distant future will be like. Pervasive uncertainty is another reason why future ethics is so challenging. If we knew *which* future would emerge, perhaps we could plan for it – focusing all our philosophical energy on resolving the questions that most matter *in that particular future*. But the possible futures are many, their comparative value is hard to discern, and each new generation will face its own new menu of possible futures. The central utilitarian ethical task is thus to enable future people to think creatively about the even-more-distant possible worlds that lie in *their* future. And we can only accomplish that task by thinking imaginatively about those futures ourselves and teaching our immediate successors to do likewise.

Uncertainty is a general problem for utilitarians. Taking the distant future seriously obviously exacerbates this problem. Our uncertainty about the future is not only empirical. It is often also normative. We can distinguish at least three dimensions of normative uncertainty: (1) we don't know what normative questions will loom largest for future people (and therefore we don't know how best to prepare them for the ethical challenges they will face), (2) we don't know what answers future people will give to those normative questions (and therefore we cannot easily predict their behaviour), and (3) we don't know what *actually* matters (and therefore we don't know whether some present choice would make things better or worse for future people). Our ignorance of future people's questions and answers exacerbates our ignorance of what the future will be like. But we also don't know whether or not any *particular* future would be *desirable*.

Utilitarians should pay particular attention to futures that are *credible*, *unsettling*, and *worrying*: futures that might happen, that destabilise the presuppositions of our current ethical thinking, and that raise very significant ethical challenges for future people.

As well as broken futures, I will explore moral challenges associated with some specific *technological* futures:

- *Virtual futures* where people have abandoned the real world altogether and spend their entire lives plugged into an experience machine that perfectly simulates any possible human experience (Nozick, 1974, pp. 42–5).
- *Digital futures* where flesh-and-blood humans are joined – or even *replaced* – by digital beings – intelligent machines and/or digital copies of human brains (see, e.g., Blackford and Broderick, 2014; Bostrom, 2014; Hauskeller, 2013, pp. 115–32; Mulgan, 2014a, 2016, forthcoming b).
- *Extraterrestrial futures* where some, most, or all future well-being is enjoyed by beings who do not inhabit the Earth. These extraterrestrial beings might be

(1) human beings spread through the solar system; inhabiting domes on Mars, the moon, or other solar bodies; travelling between the stars; or colonising distant exoplanets; (2) our distant trans-human or non-human descendants (perhaps uploaded into fully digital worlds); (3) the distant descendants of some other terrestrial species who turn out to be more robust than humans; or (4) the 'indigenous' inhabitants of some extraterrestrial environment.

Virtual, digital, and extraterrestrial futures are all credible (Mulgan, 2014a, 2018a, 2018b, 2018c). No one can reasonably assume they won't happen. Also, although they are often presented as *alternatives* to the broken future – as ways to sidestep future scarcity – these futures can also be very broken themselves. (And, crucially, as soon as we try to *evaluate* technological futures, we encounter a host of contested questions in utilitarian value theory and metaphysics.)

These possible futures are all *inhabited*. Each contains individuals whose well-being matters.[4] But we cannot simply assume that the future – especially the distant future – will be inhabited at all. Utilitarians must also consider *empty futures* devoid of creatures whose lives can go well or ill. Credible threats of human extinction make empty futures more salient than ever before. This introduces new utilitarian questions. In particular: is the *expected value* of an inhabited future positive or negative? If we must choose between empty and inhabited futures, which should we prefer?

Future ethics raises urgent new challenges. One final reason for utilitarians to focus on future ethics is the division of intellectual labour. What ultimately matters is not the importance of the question but what we can add by our contribution to it. Moral relations between affluent contemporaries have been studied extensively. While it is radical in practice, the utilitarian view of those relations is fairly familiar. It is in the new territory of future ethics that individual utilitarian thinkers (and individual books) can get the biggest bang for their philosophical buck.

2.2 Why Utilitarian Future Ethics Must Be Collective

Act utilitarianism tells us that the right act in any situation is whatever produces the best outcome. Utilitarians can depart from this orthodox position in three ways (Eggleston, 2014; Miller, 2014; Driver, 2014):

1 **Alternative foci of evaluation:** Instead of focusing on acts, we could evaluate rules, motives, dispositions, moral outlooks, social institutions, etc. We might seek the optimal rule or motive rather than the optimal act.

[4] One possible exception, as I argue below, is that it is not obvious that all digital futures are really inhabited.

2 **Indirect evaluation:** Instead of assessing acts directly, we could assess them indirectly. Perhaps the right act is whatever follows from the optimal rule or motive.
3 **Collective evaluation:** Instead of asking how each individual agent can make the world better, collective utilitarianism focuses instead on what we do together.

Most introductions to utilitarianism treat act utilitarianism as the natural default position (e.g., Bykvist, 2009; de Lazari-Radek and Singer, 2017). My approach in this Element is different. I will argue that, whatever its general merits in other contexts, individual evaluation is not the best place to begin our exploration of *future* ethics. In this section, I explore three utilitarian departures from individual direct act utilitarianism: Parfit's emphasis on what we together do, Hooker's rule utilitarianism, and my own ideal outlook utilitarianism. These introduce, respectively: a shift from individual to collective evaluation; a shift from direct evaluation of acts to indirect evaluation of acts in terms of rules, motives, or outlooks; and a focus on the moral outlook that we should teach to future generations. Each of these departures is controversial. However, they collectively offer utilitarian future ethics new and under-explored resources.

2.2.1 Parfit on What We Together Do

Our most pressing intergenerational dilemmas are large-scale, long-term collective action problems. (It makes little sense for any one individual agent to ask: 'What can *I* do about climate change?') In 1984, Derek Parfit warned that focusing on the isolated effects of individual actions can blind us to the very real and harmful impacts of 'what we together do' (Parfit, 1984, chapter 3). In cases where the impact of each individual's actions is negligible or even imperceptible, we may fail to even recognise that we are *collectively* inflicting enormous harm on future people. Perhaps the best way for *me* to promote well-being is to amass a vast fortune by emitting fossil fuels and then donate the profits to charity. If so, a consistent act utilitarian would applaud such behaviour. But if we all reason this way, the results may be disastrous. Humanity's failure to diagnose – let alone solve – the myriad collective action problems of contemporary environmentalism suggests that Parfit was right.

Parfit identifies five common 'mistakes in moral mathematics' (Parfit, 1984, p. 66). Parfit argues that these mistakes arise because we focus on the direct effects of particular actions while ignoring the systemic impact of collective patterns of behaviour. The most significant mistakes, for our present purposes, are the third, fourth, and fifth: ignoring small chances, small effects, and imperceptible effects.