2

The recurring question: why regulate utilities?

If I were asked to offer one single piece of advice to would-be regulators, on the basis of my own experience, it is that as they perform their *every single* regulatory action they ask themselves: 'Why am I doing this? Is it really necessary?'[1]

Why do we regulate the public utility industries? A number of different responses to this question have been advanced, each of which, as we shall see, tends to demonstrate both merit and limitations. The purpose of this chapter is not to conflate these distinct rationales into an all-encompassing account of why economic regulation is applied in the public utility industries, but rather to outline the different perspectives on the reasoning behind regulation of these industries.

An important distinction is drawn in this chapter between so-called normative accounts for regulation (loosely, why *should* we regulate?) and alternative accounts that attempt to explain the existence of regulation (why *do* we regulate?). Normative accounts of regulation typically focus on achieving particular aims through regulation. This approach tends to adopt particular assumptions about the application of regulation, and the ability of regulators, including that regulators: operate with good information; are able to perfectly enforce their decisions; and are generally benevolent and public-spirited in their actions. Alternative accounts of regulation typically eschew any grander purpose for regulation and focus on explaining why it is that regulation may exist in the form that it does in the public utility industries. In seeking to explain the existence of regulation, these accounts draw on economic reasoning, as well as the influence of political and legal considerations.

The question 'why regulate the public utilities?' is sometimes dealt with in a cursory way, in part it would seem because of the difficulties associated with developing a single or unified explanation for regulation.[2] Nevertheless, appreciating the different perspectives on why we regulate the public utilities is important for at least two reasons. Firstly, different rationales for regulation imply different regulatory policies and institutions, and will guide both the nature of the intervention in economic activities and markets and the form that intervention should take.[3] Secondly, any assessment of the effects of

[1] Alfred Kahn, Comment on Joskow and Noll (1981:66).
[2] As Braeutigam (1989:1299) puts it: 'In any particular case, there may be a host of possible political and economic answers to the question: Why regulate?'
[3] For the importance of distinguishing rationales for regulation more generally see Breyer (1982:34).

regulation can only proceed against the specific rationale of its imposition. For example, if regulation is predicated on a need to ensure that the public utilities set efficient prices which reflect underlying costs, any assessment should measure results against this stated objective. On the other hand, if regulation is understood as arising from the interaction of different interest groups, then we should not necessarily expect to observe prices that are lower, or that better reflect costs, as a consequence of regulation.

This chapter has two parts. The first part focuses on normative rationales for regulation: why we *should* regulate the public utilities. It begins by outlining the traditional normative argument for the regulation of public utility industries which, in general terms, is based on the close correspondence between the characteristics of these industries and the economic notion of 'natural monopoly'. While this remains the dominant specific rationale for the regulation of the public utility industries, the chapter also considers a number of other normative propositions that are used to support regulation in these industries, including those related to a need to control monopoly power or to deal with the presence of externalities in an industry. The second part of the chapter discusses various alternative accounts for the existence of regulation – that is, arguments as to why we *do* regulate. It looks first at general theoretical explanations that focus on the influence of different interest groups in regulation which have come to be known as the 'economic theories of regulation', and then at some more specific accounts for regulation in the public utility industries.

2.1 NORMATIVE RATIONALES FOR PUBLIC UTILITY REGULATION

2.1.1 Efficiency rationales for regulation

The conventional economic response to the question of why we should regulate public utilities invokes the economic notion of a 'natural monopoly', highlights the potential inefficiencies of such natural monopolies, and then draws a correspondence between the public utility industries and natural monopoly.

What, then, is a 'natural monopoly' industry structure in economics? Although the notion has changed over time, a consistent element in all accounts is that, in certain conditions, it is most cost efficient if a single firm, rather than two or more firms, produces a specific set of outputs.[4] In most cases, this situation arises where production in an industry comprises a large proportion of fixed costs (that is, costs which are incurred regardless of how many outputs are produced).

Early concepts of natural monopoly focused on the single-product case where average costs decreased as output increased for all levels of production. In these circumstances, for all levels of market demand, a single firm supplying the market would always have lower average industry costs than two or more rivals supplying different segments of the market.[5] Figure 2.1 shows long-run average costs (AC) declining for all levels of

[4] Lowry (1973) and Sharkey (1982:chp 2) provide useful discussions of the evolution of the concept of natural monopoly. Mosca (2008) provides a more recent account of the origins of the concept.

[5] Assuming that all suppliers have access to a common level of technology and face constant factor input prices.

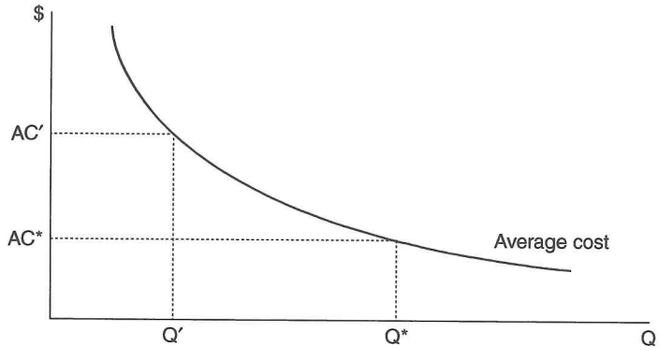

Figure 2.1 Average cost curve under economies of scale

output (Q). In this figure, industry average costs AC* at output level Q* are lower than they are at output level Q′, and indeed, at all output levels before Q* (while long-run average costs continue to fall at all output levels greater than Q*). This conception of natural monopoly focuses particularly on the characteristics associated with economies of scale in production,[6] including economies of density.[7]

Later research came to the empirical realisation that many public utility industries, in fact, supply multiple products rather than a single product. This led to an expanded notion of natural monopoly based on the recognition that economies of scope in production can, in some circumstances, also result in it being more cost effective for a single firm to supply a market. An example often referred to in this context is a telecommunications company that provides both long-distance phone calls and local calls using the same network infrastructure (e.g. copper wires from a house to an exchange). In this case it is more cost effective for a single telecommunications provider to provide both types of services rather than having two separate providers of copper wires, one for local services and one for long-distance services.

Implicit in this notion of natural monopoly, and economies of scale and scope, is an assumption about the nature of technological change in an industry. For example, the economic analysis of natural monopoly implicitly assumes that a constant, or common, type of technology is used in the production process by all firms. In this sense, the definition of natural monopoly corresponds to a given type of production technology – typically equipment that is indivisible (unfeasible to install equipment of a different size),

[6] In some work discussing natural monopoly the terms 'scale economies' and 'increasing returns to scale' are used synonymously. However, although related, the two terms are distinct. Increasing returns to scale refers to a situation where all inputs are increased by a constant amount, and this leads to a greater than proportional increase in output (i.e. if inputs increase by 2 units and output more than doubles). The concept of scale economies is broader and refers to when an expansion of output of a firm or industry results in a reduction in long-run average costs of production.

[7] Generally speaking, economies of density involve reductions in average costs associated with greater usage of a network. Examples include reductions in costs associated with more cable connections in an area, or in relation to traffic on an airline network. The implication of economies of density is that it may be efficient for networks not to overlap, and for a single firm to service a particular geographic area, or in the case of airlines, a particular route.

immobile (fixed in a specific location) and durable (expected to operate over a relatively long time scale) – the cost profile being consonant with this. The risks of this assumption are, of course, obvious in industries where technology is changing and, with it, the cost profile of production.

Two further points should be noted in relation to the notion of natural monopoly in economics. Firstly, depending on the shape of the average cost curve, it is possible for economies of scale, or economies of scope, to exist at certain levels of output but not at other levels of output. This suggests that whether an industry is a natural monopoly is conditioned by the size of the market it serves. For example, given the size of market demand in densely populated cities it may be possible that industry costs will be minimised with more than one firm. This relationship between market demand and natural monopoly is shown in Figure 2.2. In this figure, at output levels less than Q* – such as Q′ – average costs are falling rapidly and industry costs are minimised if there is only a single supplier. The minimum average cost is reached at the output level of Q*, and the industry remains a natural monopoly in both the region of declining average costs (before Q*) and where the average cost curve is flat up until 2Q*.[8] However, once the level of output reaches 2Q*, the industry is no longer a natural monopoly as two firms can each produce an output level of Q* at the same average cost as one firm can. It is also possible for economies of scope to exist at some levels of market demand, and not at others. So, for example, total costs might be minimised if only a single producer supplies two products at low levels of demand, but this condition may not hold at higher levels of market demand.

Secondly, whether an industry is defined as a natural monopoly depends on the overall production costs in that industry, having regard to economies of scale and/or economies of scope. Put differently, a natural monopoly need not exhibit scale economies (decreasing average costs) across all its levels of production, nor for all of the products it produces.[9]

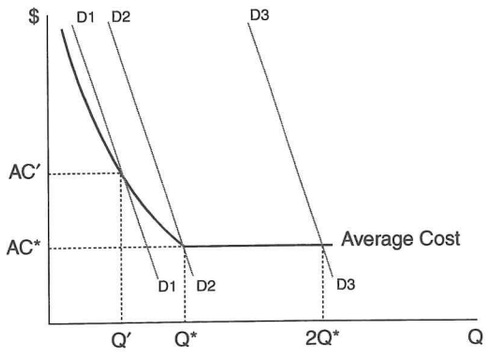

Figure 2.2 The sensitivity of natural monopoly to market demand

[8] At all points between Q* and 2Q*, given the shape of the average cost curve, a single firm still yields the least-cost production. That is, the cost function is subadditive at those levels of output.

[9] For this reason, it is often stated that economies of scale are a sufficient, but not necessary, condition for a single product natural monopoly to exist.

The critical test is that, accounting for all cost considerations, a single firm is still the most cost effective method of production. This reasoning is based on the concept of the 'subadditivity' of the cost function. In simple terms, subadditivity is said to exist where, for a given level of one or more outputs, total costs are minimised if only one firm produces these outputs rather than more than one firm, irrespective of how that output is divided among the multiple firms.[10] If we assume that q^1, q^2, …,q^m are different output vectors which sum to Q such that total output is given by $Q = \Sigma (q^1 + q^2 + \ldots q^m)$ then, assuming that all firms have an identical cost function, it would be more efficient to have a single supplier produce Q if the following condition is satisfied:

$$C(Q) < C(q^1) + C(q^2) + \ldots C(q^m) \qquad (2.1)$$

In equation (2.1), C(Q) is the total cost associated with jointly producing all of the outputs in combination, while $C(q^1)$ is the cost of producing only output q^1. If this condition is satisfied, then the cost function is subadditive which implies that it is more cost efficient for a single firm to produce the total output Q than to have the outputs (q^1, q^2, …,q^m) produced individually by different firms.

To understand subadditivity, consider first the case of a single-product industry where the cost function is such that some portion of average costs decrease as output increases up to a point, after which average costs increase as output increases. That is, the industry cost curve is U-shaped. In these circumstances, given the profile of average costs, the desired level of industry output could potentially be produced by one firm, or by more than one firm whose individual outputs combine to equal the industry output. If a single firm could supply the entire output at a lower average cost than two or more firms, even though some portion of the firm's costs are increasing in production, then this is determined to be the most cost effective production structure. In Figure 2.3, AC* is the average cost incurred at the point where demand intersects with the average cost curve at the output level of Q*. At this point the average cost of production is higher than the minimum average cost of AC′ which is attained at the output level Q′. However, although economies of scale only exist up to point Q′, and diseconomies exist thereafter, because the cost function is assumed to be subadditive, it is still most efficient for a single firm to produce the output level of Q*.[11] Figure 2.3 therefore demonstrates the central point that a subadditive cost function does *not* require economies of scale to be present over the entire relevant range of output (in this case, the output up to Q*).

The same reasoning applies when a production process involves multiple products or services. In these circumstances, it is necessary to consider the cost conditions associated with all of the different outputs provided by the firm (i.e. long-distance calls, local calls, etc.) when considering whether the industry displays the characteristics of a natural

[10] Strict subadditivity is defined to be where 'the cost of the sum of any m output vectors is less than the sum of the costs of producing them separately'. See Baumol (1977:809).

[11] If a second firm entered to produce only Q*–Q′ then, on the basis of the assumption that costs are identical across the industry, this firm would incur a very high level of average costs on that small level of output, which would raise the overall industry costs above that which would exist if a single firm supplied all of the output to Q*.

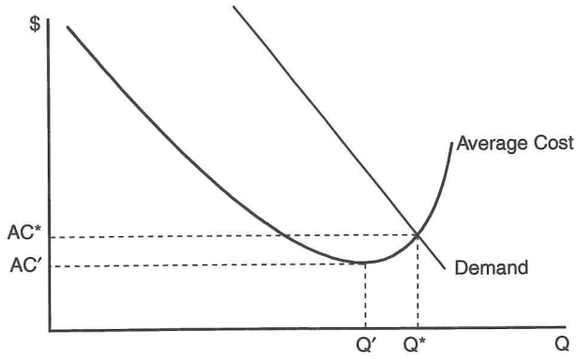

Figure 2.3 Subadditivity without economies of scale over the entire relevant range of output

monopoly. Again if, having regard to the costs associated with the different products produced, the cumulative average costs of one firm producing all products is lower than the cumulative costs associated with structures with two or more firms supplying different combinations of the products, this satisfies the condition for a natural monopoly. It follows from these points, that the standard definition of a natural monopoly industry commonly employed in regulatory discourse today, is an industry structure where, over the relevant range of market demand, the cost function of a firm is subadditive.

Correspondence with public utility industries

Public utility industries – or, more accurately some parts of public utility industries – are generally considered to have attributes that make it most cost efficient if a single firm rather than multiple firms produces a specific set of outputs.[12] The provision of public utility services typically involves investments in durable and immobile assets, such as electricity wires and poles, copper wire or fibre-optic cables, gas pipelines, or water and wastewater pipes. More specifically, the costs of production typically involve a large fixed component, which are sunk once incurred, and then low or negligible on-going operational costs associated with the production of each unit. For example, electric, telecommunications and gas transmission and transportation networks involve large and lumpy capital investments of a long-term nature, that are sunk once incurred, and are followed by relatively low variable or marginal production costs. This gives rise to a cost profile in which average costs decline as production increases. That is, as production increases, the high level of fixed costs (which by definition are largely invariant to levels of output) can be spread across a greater number of output units; it is this ability of a firm to spread such costs over a large level of production that reduces the level of long-run average costs. If the fixed costs associated with the construction of a gas pipeline are $100 million, and there are 100,000 users, the attributed fixed cost per user is $1,000, while if there are 1 million users of the pipeline the attributed fixed cost per user is $100.

[12] Newbery (1999:27) describes network utilities as 'the clearest example of natural monopolies'. See also Kahn (1971:2) and Scherer (1980: 482).

Having said this, as discussed in Chapter 1, the perceived 'natural monopoly' activities of the public utility industries has changed in many jurisdictions in recent years. In those jurisdictions where restructuring policies have been implemented, the supply chain for public utility industries now comprises a mix of natural monopoly-like activities (the core network activities) and potentially competitive activities.

Two types of efficiency arguments are generally posited for regulation of certain public utility activities on the basis of a natural monopoly structure. The first concerns allocative efficiency, and here regulation of prices is predicated on a desire to maximise total surplus and economic welfare. The second argument relates to issues of productive efficiency, and here the principal rationale for regulation is to control entry in the industry so as to avoid the wasteful duplication of fixed costs, or to avoid entry by firms who offer no new products or productive technologies and enter the market only to service a select group of the most profitable customers (so-called 'cream-skimming'). It follows that the standard regulatory prescription, where an industry structure resembles that of a natural monopoly, is to restrict entry to only one firm and, at the same time, to impose price regulation on that firm to set efficient prices which maximise economic welfare (we consider various alternatives to this standard prescription in Chapter 3).

Price regulation to achieve allocative efficiency

An important conventional economic rationale for the price regulation of public utility industries is to address concerns about allocative efficiency. In very general terms, allocative efficiency refers to an 'allocation' of products such that the marginal benefit that consumers obtain from consuming an additional unit of the output (as represented by the demand curve) is equal to the marginal cost of producing that additional unit of output. The measure of relative allocative efficiency is typically defined in terms of the concept of 'total surplus': the sum of the consumer surplus and producers' profit for a given level of production. Total surplus can be measured as the monetary difference between the benefits of consumption of a service *less* the costs of producing the service or, in more technical terms, the area below the demand curve but above the marginal cost curve for a given level of output.

According to standard microeconomic reasoning, the standard profit maximising condition for a monopoly involves producing a level of output which equates its marginal revenue to its marginal cost. Assuming that the monopolist faces a downward sloping linear demand curve, this will imply that the level of output is lower, and prices higher, than that of a competitive market where prices are, in theory, set at (or close to) marginal costs (and allocative efficiency therefore obtained). This is illustrated in Figure 2.4, where assuming constant marginal costs, P^* and Q^* is the price and output associated with perfect competition, and P^{Mon} and Q^{Mon} are the price and output associated with monopoly. The deviation between the level of output (prices) in a monopolistic industry and that in a competitive industry is referred to as the allocative inefficiency of monopoly and gives rise to what is known as a 'deadweight loss' (roughly, as consumers pay more than it costs to produce the last unit of output, this means that there are unrealised gains from trade which could be realised or, put differently if the price was closer to marginal cost more consumers would purchase the product).

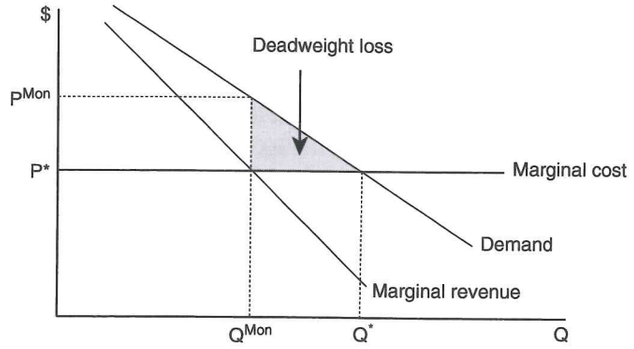

Figure 2.4 The allocative inefficiency of monopoly

The most allocatively efficient outcome (sometimes referred to as the 'optimal' outcome) is one where the total surplus is maximised, that is, when the area between the demand curve and the supply/marginal cost curve is greatest for a given level of demand and marginal cost.[13] Standard microeconomic reasoning suggests that total surplus will be maximised by selecting a level of output where prices are set to marginal cost, and where firms will use the least-cost inputs in the production process. This is illustrated as the shaded area in Figure 2.5, and is the point at which the area under the demand curve, but above the marginal cost curve, is greatest. As this is the maximum total surplus that can be obtained, regulation which imposes a pricing policy that induces such a position is known as 'first-best', as each extra unit of output produced is equal to consumers' willingness to pay for it (simply, price equals marginal cost).

However, as already noted, production in public utility industries typically comprises a large proportion of fixed costs, and in these circumstances regulatory policy makers can face a dilemma when attempting to obtain the first-best outcome described above. Specifically, if the firm is compelled by regulation to set a single uniform price for all of its output which equals the marginal cost (in order to maximise allocative efficiency), this will be unsustainable over the long term as the firm will not be recovering any of its fixed costs of production. Figure 2.6 illustrates the losses that can potentially arise if public utility firms with a high proportion of fixed costs are required to set 'first-best' prices (i.e. price equal to marginal cost). In this figure, the shaded area represents the losses incurred by the firm if it is required to set prices equal to marginal cost (P^0) rather than at average cost (P'), which would allow it to recover both its fixed and marginal costs for a given level of output Q^0. In Chapter 4, we consider a number of possible solutions to this dilemma, including the use of a subsidy to cover the fixed costs of production (equal to the shaded

[13] In competitive markets, where all firms are price takers, the supply curve is the sum of the marginal cost curves for the individual firms. The analysis here of total surplus ignores the relative distribution of total surplus between consumers and producers. As we see in Chapter 4, if an unregulated monopolist engages in price discrimination this can maximise total surplus, but the producer obtains the entire surplus and there is no consumer surplus.

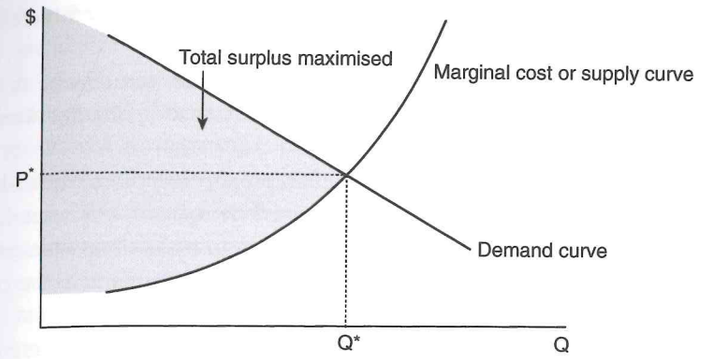

Figure 2.5 Total surplus maximised: 'first-best' pricing

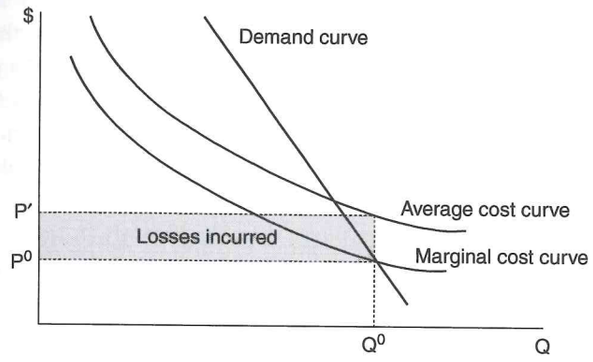

Figure 2.6 Losses which can arise from 'first-best' pricing

area in Figure 2.6), or the use of different (non-linear) pricing policies that allow the firm to recover its fixed costs from different customers or products.

While the need for price regulation to address allocative inefficiency is a conventional explanation for regulation of the public utility industries (and features in the first pages of almost all textbooks) there is, as discussed below and in Chapter 4, a large body of empirical evidence which suggests, that, in practice, regulation in the public utility industries has not been focused on implementing pricing structures designed to achieve allocative efficiency. For example, the use of demand-reflective prices (so-called Ramsey–Boiteux prices), peak-load prices or other forms of non-linear pricing which can improve allocative efficiency, have generally not been adopted by regulators principally, it appears, because of information limitations and distributional concerns.

Entry regulation to achieve productive efficiency

A different efficiency argument for regulation of industries which have natural monopoly characteristics, but where it is possible for entry to occur, is based on the proposition that it may be productively efficient (result in lower total costs, and consequently lower

average prices) for only a single supplier to provide relevant services, and therefore for entry to be restricted.

There are a number of different dimensions to the argument that regulation should restrict entry into industries with natural monopoly characteristics. The first dimension relates to the potential inefficiencies associated with duplication of fixed costs where rival firms compete. Given the cost profile of natural monopolistic industries, allowing competition in these industries will reduce productive efficiency, as it will necessitate the recovery of two or more sets of fixed costs of production. In these circumstances, entry regulation may assist productive efficiency by avoiding a situation where the costs of production are duplicated across the industry.

An older, and more controversial, rationale for the regulation of entry is that competition in activities that have natural monopoly characteristics can, in some circumstances, be 'destructive' and lead to price volatility and instability in the industry.[14] 'Destructive competition', it has been argued, has the potential to emerge in conditions where the proportion of fixed sunk costs is large as a proportion of total costs, where there are substantial periods of excess capacity, and where marginal costs lie below average costs for substantial periods of time (as in the standard definition of natural monopoly).[15] In these conditions, if entry is feasible in relation to all or some of the production activities of a firm then, as capacity becomes tight, this may, assuming relatively inelastic demand, lead to large increases in prices for an extended period before new capacity can be developed or put into service. The prospect of earning prices significantly in excess of cost may encourage entry, including the building of new capacity, to exploit the conditions of tight supply. However, once all the new capacity enters the system, and the industry is again in a situation of excess capacity, prices will tend toward marginal cost as a result of intense competition, resulting in bankruptcies. Over time, competition in these conditions is seen to create a situation of instability both in terms of consumer prices and producer profits. In such contexts, regulation in the form of restrictions on entry is argued to promote stability, and protect consumers and businesses from the effects of this intense and destructive competition.[16]

Finally, it has been argued that, in some circumstances, entry into natural monopolistic industries by rival firms, who offer no new products or production efficiencies, may be socially inefficient.[17] In particular, where there are common costs of production relating to large and indivisible investments, and which must be recovered across all customers, potential entrants, who offer no new products or productive technologies, may be encouraged to enter the market to provide only the most profitable services or service only the most profitable customers. This potential has been recognised in research relating to the 'sustainability' of natural monopoly, and whether there exists a set of so-called

[14] See Ely (1937) as referenced in Sharkey (1982:16).
[15] See Kahn (1971:173).
[16] See Sharkey (1982:16). Helm and Jenkinson (1997:1) note that the monopolistic structure of the utility industries in the post-war period in the UK was seen to prevent 'the destructive competition which was widely thought to have pervaded the industries in the 1920s and 1930s'.
[17] See Faulhaber (1975) and Panzar and Willig (1977).

'sustainable prices' for the service(s) provided by a naturally monopolistic firm.[18] In the standard paradigmatic natural monopoly case of a declining average cost curve for all levels of production, then it will always be possible for the firm to deter entry by rivals by charging a price where the average cost curve intersects with the demand curve. However, where the average cost curve is not monotonically decreasing for all levels of production then, even in the single-product case, it is possible that rival firms may profitably enter and serve only that section of the market where price is greater than average cost.[19] This practice of selective entry at the margin is sometimes referred to as 'cream-skimming' as entrants are seen to capture the 'cream' services, leaving the unprofitable services to the more established firm.[20] However, because this type of selective entry can make it unprofitable for an incumbent natural monopoly firm to supply the rest of the market, regulation has been seen as necessary to protect the natural monopoly from entry of this type.[21]

Despite the various economic rationales for restricting entry into some public utility activities described above, this form of regulation is controversial. Many economists advocate caution in restricting entry into the public utility industries.[22] In particular, automatic entry restrictions have been argued to be unnecessary under certain theoretical[23] and real-world[24] conditions. Moreover, restrictions on entry are often argued to hinder dynamic efficiency improvements in an industry by insulating the 'protected firm' from market pressures to adopt new technologies or cost-reduction techniques.

2.1.2 Regulation to control monopoly power

A second set of normative arguments for the economic regulation of the public utility industries is based on the fact that in many cases – either as a result of restrictions on entry,

[18] A sustainable price vector is one which offers no profitable opportunities for entry for firms who offer the same service(s) and face the same cost functions as the natural monopoly, but allows the natural monopoly to satisfy all demand and to break even across the portfolio of products it supplies (i.e. to earn a normal profit).
[19] Panzar and Willig (1977:1) examine the concept of sustainability of natural monopoly in the context of the production of multiple products (where rival firms may enter and seek to specialise in the supply of only one or more of the services provided by the natural monopoly) and conclude that strong demand substitution effects and product-specific scale economies work against sustainability. Similarly, Faulhaber (1975:974) concludes that where average costs are U-shaped and entry is free and costless, there may be no stable supply arrangement for a natural monopoly. This is because for any set of prices for a particular coalition of customers, there will always be an incentive for a firm to enter the market and offer lower prices to supply a different sub-set or coalition of customers.
[20] See Viscusi, Harrington and Vernon (2005:533).
[21] As described in Chapter 10, the 'duopoly' policy that applied in the telecommunications industry in some jurisdictions in the 1980s can be seen as a practical example of such entry restrictions. In the UK the entrant (Mercury) was limited to a market share of voice telephony of 3 per cent of that of BT which, according to Spiller and Vogelsang (1994:21), 'was meant as a safeguard for BT's envisaged large investment program in expansion and modernisation of its network'.
[22] See Panzar and Willig (1977:21); Joskow and Noll (1981:16); Vickers and Yarrow (1988:58).
[23] See the discussion of free entry and contestable markets in Sharkey (1982: chp 7).
[24] Armstrong, Cowan and Vickers (1994:106), for example, acknowledge the logical possibility of selective entry, but doubt it provides a 'good case for entry restrictions in the utility industries, which are not for the most part remotely contestable and where there is little evidence that cost conditions give rise to non-sustainability'. Similarly, Carlton and Perloff (2000:668) note 'Although it is theoretically possible that natural monopolies are unsustainable, there is little empirical evidence in most regulated industries showing that sustainability problems might justify regulators forbidding entry'.

or because of cost or technological reasons – there is only a single supplier of a public utility service, and therefore this operator may have an incentive, and the ability, to behave in ways that exploit its position of power. A monopoly provider might, for example, set prices considerably above underlying costs, degrade quality, or be insufficiently responsive to cost and other production efficiencies. To address this potential, regulation in the form of price controls and other behavioural regulations (relating to quality, etc.) are argued to be necessary.

This rationale for regulation is distinct from the one described in section 2.1.1 which focused on the fact that firms in a naturally monopolistic position do not have a natural incentive to set efficient prices which maximise economic welfare. The argument for regulation to control monopoly power is not primarily based on a desire to achieve allocatively efficient prices (a specific outcome), but rather on controlling the conduct of the firm so it does not harm consumers, either through charging prices which deviate to a considerable extent from the underlying costs of the activity, or from degrading quality or failing to invest, etc.[25] In effect, while traditional arguments for regulation have generally been framed in terms of efficiency concerns, this rationale for regulation also incorporates equity considerations. In essence, these relate to concerns associated with 'unjust' or 'exploitative' pricing, and to the ability of firms in a monopoly position to exploit their position of market dominance. This rationale for regulation may explain why we observe some activities in the public utility industries which do not have natural monopoly attributes sometimes being subject to price regulation (such as electricity generation, mobile telephony, retail supply, as well as certain transmission activities where some competition exists e.g. certain gas pipelines, and some trunk networks in fixed telecommunications networks).

Regulation premised on the need to prevent abuses of power is consistent with perspectives that regulation is a response to the fear of 'hold-up'. In economics, 'hold-up' arises in situations where either a public utility firm, or its customers, make sunk investments or incur expenditure on the basis of expectations of the future conduct of the other party, and that other party then acts opportunistically and exploits this fact after the investments have been made. Public utility companies frequently incur significant capital costs when making long-term and immobile investments (such as building transmission or distribution networks) and this is premised on an expectation that demand for these services will continue, and that the returns will allow the company to cover its costs. At the same time, users of public utility services (such as consumers) also make decisions on the basis of expectations regarding the future conduct of the natural monopoly firm – for example, in deciding which type of energy source to use for its activities (gas or electricity).[26] Regulation, in this context, is seen as a method of protecting

[25] This reasoning seems consistent with the perceptions of some regulators as to their wider legitimacy. For example, the Chairman of the Australian Competition and Consumer Commission (ACCC) (which is responsible for public utility regulation) has observed that: 'The current rationale given by most economists … is that we regulate for reasons of allocative efficiency, or to reduce dead weight loss … Most Australians would, of course, be surprised by this. They think we regulate to make sure that the owners of monopoly infrastructure do not take advantage of their position and "gouge" consumers' (ACCC 2012a: 2).

[26] See Goldberg (1976:433).

both parties against opportunistic behaviour by the other party once they have made sunk investments.[27]

Regulation in the face of monopoly power is also premised on cost efficiency arguments. Specifically, it is argued that, because the monopolist does not face the threat of competition, it will produce at higher levels of cost than firms who operate in competitive markets (who are naturally incentivised to cut costs to improve profitability and remain competitive). There are at least three potential causes of this cost inefficiency. The first cause is technical inefficiency, which relates to how various inputs (such as capital and labour) are combined in the production process to create the outputs of the firm. The general argument here is that a monopolistic firm may not have the appropriate incentives to ensure that the rate of conversion of inputs to outputs is at its most efficient and, in technical terms, that the firm sits on the efficient frontier of production (which as we will see in Chapter 5 is a major driver behind certain forms of price control arrangements). A second potential cause of cost inefficiency is that the management of a natural monopoly will face lower incentives to seek out cost savings, or to take risks, than managers in competitive markets. In the oft-quoted words of John Hicks, 'the best of all monopoly profits is a quiet life'.[28] In this respect, it is assumed that higher levels of managerial activity will generally be associated with lower costs. This cause of inefficiency, which has been termed 'x-inefficiency' has been distinguished from technical efficiency, and is intended to capture the more general case where 'for a variety of reasons people and organisations normally work neither as hard nor effectively as they could'.[29] X-inefficiency is considered likely to be greater in situations where competitive pressure is severely limited, such as monopoly, and is argued by some to be more detrimental to welfare than allocative inefficiency.[30]

The third potential cause of cost inefficiency, which is of particular relevance in many public utility industries, is how monopoly power affects the incentives of firms to innovate and to seek out and employ techniques, efficiencies and working practices in ways which could lead to future improvements in economic welfare. This issue is a difficult and controversial one in economics; however, the main arguments can be stated simply here. On one side, firms in most competitive market structures are argued to have a natural incentive to invest some resources to innovate or develop new working practices, as any successful innovations could result in the firm gaining a competitive advantage over its rivals and increase its profits above the competitive level. In contrast, in monopoly structures, these same incentives will not apply and the monopoly firm will only innovate where the expected additional profit associated with the innovation is greater than the resource costs that need to be applied (this is because the profits of the monopolist are already potentially above the competitive level). There is, however, an alternative perspective: that the ability to occupy a monopoly position, and to reap monopoly profits, can itself act as an important spur to innovation over the long term, and that this can

[27] See the discussion in Williamson (1976:91), and more recently in Biggar (2009).

[28] Hicks (1935:8). Before that, Adam Smith (1776:148) described monopoly as 'a great enemy to good management'.

[29] See Leibenstein (1966:413).

[30] Leibenstein (1966:395, 399). Contrast with Stigler (1976:213).

create an incentive for firms to invest in research and development.[31] Empirically, there is some evidence that the relationship between product market competition and innovation is non-linear, and more specifically that an inverted-U relationship holds, with industries distributed across both the increasing and decreasing sections of the U-shape.[32]

The control of monopoly power as a rationale for regulation has strong explanatory power in terms of understanding both why the public utility industries are subject to regulation and why, in practice, we do not always observe regulation operating in ways consistent with the natural monopoly rationale outlined above. However, it does leave some questions unanswered. It does not explain, for example, why firms who hold close-to monopoly positions in non-public utility industries are not subject to the same forms of economic regulation (i.e. regulation over and above normal competition law) as the public utility firms. It also does not explain why the regulation of the public utility industries generally takes the form of *ex ante* controls (on prices, quality, etc.) rather than *ex post* controls (such as competition law prosecutions for suspected abuses of monopoly power). This issue is discussed in Chapter 3.

2.1.3 Externalities as a rationale for regulation

A further normative rationale for economic regulation of the public utility industries is the need to address the externalities that frequently arise in these industries. In general terms, externalities arise where there are wider costs or benefits associated with the supply of a service than those that accrue to the immediate parties to the transaction (i.e. there are uncompensated third-party effects). Externalities take many forms, and there are both positive and negative externalities. The most familiar type of negative externality associated with the public utility industries – particularly the electricity and wastewater industries – relates to pollution, and harm to the environment, that may be associated with how services are produced and supplied. Examples of positive externalities include the widespread benefits associated with the provision of clean drinking water and adequate sanitation (which reduces the spread and cost of illnesses) or extensive transportation and communications networks (which allow more people to connect with one another). In each of these cases, regulation can be premised on the need to ensure that the wider societal benefits/harms of transactions in certain services are realised/avoided.

The presence of significant externalities as a rationale for regulation of the public utility industries is not a new one.[33] It has, however, become increasingly relevant in many jurisdictions in the context of changes associated with the environment. In the energy, water and transport industries (and, to a lesser extent, communications) regulators in some jurisdictions now see an important role for themselves in representing or protecting the views of future consumers/citizens when considering the impacts of

[31] The most important reference here is that of Schumpeter (1943:81). The high levels of innovation in pharmaceutical markets, where firms have exclusive set period monopolies over the supply of new and innovative drugs, are an example often noted.
[32] See Aghion, Bloom, Blundell, Griffith and Howitt (2005:701).
[33] See Kahn (1971:236).

current decisions and policies. In effect, such regulators are attempting to identify, and address, the externalities associated with the current practices of public utilities on future generations.

A particular type of positive externality in the public utility industries are network externalities.[34] Network externalities arise where the benefits to one user of a network depend on the number of other users that are connected to, or utilise, the network, and is a characteristic typically associated with communications and transportation networks in particular. The essential argument for regulation in these circumstances centres on the realisation of the benefits that can collectively arise to all users of a public utility network when there are a large number of other users who also use the same network. The most obvious example is that of a telephone network, where each owner of a telephone benefits when the size of the telecommunications network increases and they can contact greater numbers of other telephone users.[35] In short, as a network becomes larger, the value of the network to each and every consumer/user of the network increases (i.e. each additional user connected to the network inadvertently creates a benefit to existing users).

Economic regulation can harness network externalities in two ways. Firstly, regulation to restrict entry can allow one firm to internalise the benefits associated with a larger network The focus of regulation here is on improving welfare by ensuring that consumers benefit from a large network using a common technology (compared with smaller competitive networks with different compatibility standards) rather than on the productive efficiency of the single supplier (as in the natural monopoly discussion above). A second way in which economic regulation may address network externalities is through price regulation. Specifically, the prices charged for network use (or to specific categories of user) can be adjusted from the underlying cost to account for the benefits associated with network growth and a larger network. As discussed in Chapter 10, an example of this approach can be seen in the mobile telephony industry in some jurisdictions where regulated prices at the wholesale interconnection level have sometimes been increased (in the form of a network externality surcharge) to encourage mobile phone network operators to reduce retail subscription prices, and thus increase mobile phone penetration.

While a normative rationale for regulation based on the existence of externalities has been around for many years, and is widely accepted, some are circumspect about intervention on this basis. In particular, it has been argued that the mere identification of an externality should not automatically justify regulation to subsidise the development of a network.[36] Instead careful judgement should be applied in determining which externalities require regulatory intervention, and the forms that intervention should take.[37]

[34] A distinction is sometimes drawn between the definition of network externalities and network effects. See Liebowitz and Margolis (1994:135).
[35] See Rohlfs (1974).
[36] Kahn (1971:195) argues that a public policy decision to subsidise provision of electricity or telephone services to particular sectors of the populace is, in principle, no different from the decision to provide them with a decent diet, medical care and housing, and that all of these services should be provided by devices other than regulation.
[37] See, generally, Coase (1960).

2.2 ALTERNATIVE EXPLANATIONS FOR PUBLIC UTILITY REGULATION

The various normative rationales for regulation discussed in section 2.1 imply that regulation of the public utility industries is generally premised on a desire to improve economic welfare, either by requiring firms to act efficiently, ensuring an efficient industry structure (through restricting entry), preventing abuses of monopoly power, or addressing various economic externalities. In effect, each rationale is tied to the generation of positive economic welfare improvements.

However, there is considerable empirical evidence suggesting that the practice of regulation is not always consistent with a goal of improving economic welfare.[38] Studies and surveys have suggested, for example, that the effects of regulation can differ considerably from what the public interest theories suggest.[39] Moreover, there are a range of activities in the economy that have historically been subject to some form of economic regulation, yet do not appear to be either natural monopolies, or feature characteristics consistent with the other economic rationales for regulation discussed in Section 2.1.[40]

For these reasons, it has been argued that a richer, multi-dimensional account is needed to explain why regulation exists in the public utility industries and the form its takes. In the discussion that follows we consider some of these alternative explanations. One set of explanations, known collectively as 'the economic theories' or 'interest group theories' of regulation, suggests that regulation is best explained by considering the different political and economic actors who interact in society and their incentives: including politicians or bureaucrats; regulated companies; consumers and other powerful interest groups affected by regulation, such as workers in a regulated industry. Another explanation conceives of public utility regulation as a response to the need for some form of administrative arrangement to manage the long-term relationship between consumers and producers of public utility services. Other explanations point to the political and social importance of utility services, and suggest that public utility regulation exists, and takes the form that it does, in part, to address various distributional issues, including issues relating to fairness and the protection of various groups in society.

2.2.1 Economic or interest group 'theories' of regulation

An important set of articles published in the 1960s and early 1970s directly challenged the notion that regulation exists solely as a mechanism to address normative concerns about natural monopoly and to improve economic welfare (and, in particular, to compress the gap between price and marginal costs that would otherwise exist in these industries).[41]

[38] Joskow and Noll (1981:36) find that, when considered as positive theory (rather than normative guidance), the public interest theories of regulation are wrong, being generally inconsistent with available evidence.

[39] The principal references here are: Jarrell (1978:276); Noll (1989:1254); Joskow and Rose (1989:1496); Knittel, (2006:203); Biggar (2009).

[40] See Posner (1974:336) and Stigler (1971:4).

[41] This work, and particularly George Stigler's 1971 paper, has been described as 'the beginning of the end' of the widely held assumption that regulation was introduced to pursue widely accepted public interest goals. See Joskow (2005a:189); Peltzman (1993:822).

This work argued that the existence of regulation was more accurately accounted for by a propensity of different groups in society to demand, and then utilise, regulation to achieve private gains and benefits. According to this reasoning, regulation exists not to protect the interests of the public at large, but to represent and protect the interests of specific politically effective groups. Collectively, these theories of regulation are sometimes referred to as the 'economic theories of regulation' or 'interest group theories', although there are a number of different strands within this work, particularly in relation to whom regulation is intended to serve (producer interests or other interest groups), and the mechanisms by which different groups in society are able to influence regulatory outcomes.

The first strand of these theories builds on the proposition that the existence of regulation might be explained by a desire of firms themselves to be regulated.[42] Evidence of the introduction of state-based regulation of the electric utilities in the USA, and telecommunications regulation, in the early twentieth century appears to support this view.[43] Regulation, which was viewed primarily as a pro-producer policy, was in greatest demand by utilities operating in competitive market conditions with low prices and profits.[44] The potential benefits to public utility firms of regulation were argued to include: direct subsidies to the industry; control over entry by new rivals; and actions to promote complements and restrict substitute products.[45]

Intuitively, it might be argued that regulation would, despite the above, be unattractive to producers insofar as it may require that prices be set to reflect costs and ensure that regulated firms earn only a normal profit. However, empirical studies suggest that, in practice, regulation does not necessarily result in reductions in prices, improvements in efficiency or reductions in industry profits.[46] Nor does regulation necessarily protect consumers from the exploitation of monopoly power by utilities.[47]

[42] See Hayek (1944:48), for an early exposition of the proposition that aspiring monopolists regularly sought, and frequently obtained, the assistance and power of the state to make their control effective.

[43] See Gray (1940:9) who argued that the introduction of state-supported public utility regulation in the electric utilities in the USA in the early twentieth century provided a 'haven of refuge for all aspiring monopolists who found it too difficult, too costly, or too precarious to secure and maintain monopoly by private action alone'. Later, Brock (2002:52) describes the efforts of AT&T chairman, Theodore Vail, in the early part of the twentieth century to 'embrace regulation and use it is as substitute for market forces', noting that 'Vail recognised that regulation could be a way of preserving monopoly power in justifying a system without competition'. See also Demsetz (1968:65), who argues that regulation provided utilities with 'the comfort of legally protected market areas', and that the force behind the regulatory movement came from the utility companies themselves.

[44] See Jarrell (1978:293).

[45] See Stigler (1971:5). However, these benefits would come at a cost to the industry, which would take the form of votes and resources. Stigler notes that resources might take the form of campaign contributions, contributed services (a businessman heading a fund-raising committee) or indirect methods (such as the employment of party workers).

[46] Jarrell (1978:293) studied the effects of regulation on pricing in the US electricity industry, for example, and concluded that prices and profits increased upon the establishment of state regulation. Stigler and Friedland (1962:11) find only a very small, and statistically insignificant, effect of regulation on electric utility prices. However, as Peltzman (1993:820) notes, there were various mistakes with the original Stigler and Friedland statistical model (such as a coding error on the dummy variable), which mean that the original results are wrong about the magnitude (but not the statistical significance) of the effect of regulation. Joskow and Rose (1989:1466) highlight other reasons for exercising care when generalising these results.

[47] See Jordan (1972:163). Contrast with Joskow and Rose (1989:1496) who conclude, based on their survey, that price regulation does reduce prices below those which an unconstrained monopolist with an exclusive franchise would choose, but that the structure of prices and distribution of revenues often reflect distributional objectives rather than efficiency objectives.

Building on work on collective behaviour and clubs,[48] Stigler sought to explain the supply and demand for regulation by groups in society more fully, developing the idea that policy makers and regulators might act as rational actors when confronted with the political demands of interest groups (particularly producers). Stigler's central conclusion is that more highly organised groups (which are typically smaller in size), and who have large stakes in the outcome, will generally be more successful in acquiring regulation for an industry. It follows from this reasoning that producers or sellers in a particular industry – who are generally smaller in number than consumers, and have higher *per capita* stakes in the outcome of regulation – would be expected to be relatively more successful in bidding for the services of regulation than consumers.[49] Stigler's conclusion is often seen to be consistent with the more general 'capture theory' of regulation, whereby regulation is seen to exist to serve, or be applied in ways consistent with, the preferences of the incumbent regulated firm.[50]

However, in work published around the same time, Richard Posner, found no single interest group responsible for the capture of regulation. Specifically, Posner argued that the prevalence of cross-subsidisation in regulated industries could not be explained by reference to the view that regulation was pro-producer insofar as the regulated firm would always be better to stop supplying the below-cost service than subsidising it from other activities.[51] Posner concluded from this that regulation could perform allocative and distributive functions that were normally associated with taxation by government. Posner's analysis broadened the interest group approach view of regulation by suggesting that certain groups of customers may also have a demand for, and an effective influence on, regulation.[52]

Two subsequent papers expanded on this theme that regulation serves a broad constituency.[53] Peltzman (1976) examined how regulation affects the transfer of wealth among different interest groups, where the regulator is subject to some form of regulatory budget constraint. According to this approach, a regulator seeks to make everyone who has political weight 'as happy as possible' and to obtain a politically optimum distribution of wealth, as reflected in profits that producers can earn and the prices charged to consumers.[54] This framework is potentially able to simultaneously explain the existence of both pro-producer-type regulatory outcomes, as well as outcomes such as the existence of cross-subsidisation of some services in regulated industries. In each case, the resulting

[48] Particularly the work of Olson (1965) in relation to the size of groups and the incentives to affect political outcomes, and Buchannan (1965) on the theory of clubs.
[49] See Peltzman (1989:8).
[50] The origins of the proposition that public policies might reflect competition among different interest groups can be found in the works of Bentley (1908). See also Bernstein (1955:chp 3) on the susceptibility of regulatory agencies over time to capture by the regulated industry.
[51] Posner (1971:27). To Posner, the existence of cross-subsidisation was an 'embarrassment' to those who saw regulation as being imposed to bring about results approximating a competitive market, as it resulted in an outcome 'unthinkable in a competitive market' (i.e. prices for some services below cost).
[52] Posner's analysis is not inconsistent with Stigler's; Stigler's analysis allows for the capture of regulatory processes by other effective political groups, not just regulated firms.
[53] Joskow and Rose (1989:1497) conclude, on the basis of their survey, that labour, in particular, can be an important beneficiary of regulation in certain industries, and even more so than regulated firms. They argue that price and entry regulation is conducive to the development of strong unions.
[54] Peltzman (1989:10).

equilibrium reflects a balance between political considerations, such as the weight and influence of different interest groups, and economic components, such as the demand and cost conditions in the industry.[55]

Adopting a similar framework, Becker (1983) developed the proposition that regulation involves the balancing of considerations of redistribution and efficiency, and concludes that the resulting equilibrium will be determined by the size of the deadweight loss which results from the inefficiency of regulatory policies.[56] An important implication of this analysis is that regulatory policies directed at efficiency will be successful in circumstances where the relative gains to those who favour such policies is much greater than the relative losses of those who oppose them.[57] This implies that one reason why we see regulation of the public utility industries is that there are significant market failures, and the potential exists for significant efficiency gains as a result of the introduction of regulation.[58] Becker concludes that this analysis unifies the different views on regulation: that it can correct for market failures (such as the inefficiencies associated with natural monopoly), while at the same time favouring politically powerful groups.

Collectively these economic theories, or interest group theories, of regulation, are seen to have advanced the understanding of economic regulation, particularly the political context in which regulation occurs.[59] However, over time, a number of criticisms and critiques of these theories have emerged.[60] In particular, it has been argued that the 'economic' or 'interest group' theories of regulation, like the normative theories of regulation they challenge, are effectively generalisations rather than 'theory' and have not been systematically confronted with wide-scale empirical testing.[61]

At a more analytical level, it has been argued that the theories fail to account for various information asymmetries (between regulated firms and regulators, and between regulators and oversight bodies), and to distinguish between political and regulatory institutions, including taking account of the agency relationship between the government/the legislator and regulatory agencies.[62] It has also been argued that legislators and

[55] An equilibrium in this framework will be where the marginal benefit in terms of votes gained by raising profits for the regulated firm is exactly offset by the marginal loss in terms of votes lost resulting from an increase in prices for consumers; with the consequence that the resulting price will lie somewhere between the profit maximising price (suggested by the pro-producer theory) and the perfectly competitive price with zero profits (suggested by the public interest theories).
[56] Deadweight loss in this context is the difference between the gain to the winner of political influence less the loser's loss from a change in output which can be attributed to regulation. See Peltzman (1989:12)
[57] Becker (1983:396).
[58] See Viscusi, Harrington and Vernon (2005:388).
[59] Stigler's (1971) paper, in particular, is seen to have produced a significant shift in the 'professional center of gravity towards a skepticism [among economists] about the social utility of regulation', see Peltzman (1993:824).
[60] Peltzman (1989) presents an excellent survey of these critiques.
[61] See an early critique by Posner (1974:352). There is a recognised difficulty in measuring and causally testing the variables that comprise the theory (such as the relationship between the stakes of a particular group and the gains it receives, or which interest groups will be successful), making the rejection of the null hypothesis virtually impossible. See Joskow and Noll (1981:39); Noll (1989:1277). However, Knittel (2006) finds empirical support for the interest group theory of regulation. See also Ando and Palmer (1998).
[62] See Laffont and Tirole (1991:1090).

regulators may have particular ideological concerns which may, in some circumstances, override any obligations that they feel they have to particular interest groups. More generally, even some leading proponents of the economic theories of regulation accept that such theories are unable to provide a coherent account of some important questions about regulation, including why regulation only applies to specific industries, why it is introduced when it is, and why deregulation has occurred in some industries but not others.[63] Nevertheless, the conceptual paradigm, which combines economics and politics, and highlights the potential susceptibility of regulation to organised interests, has provided an important and enduring framework for considering the role of different interests and influences on the existence and conduct of regulation.

2.2.2 Regulation as a form of administration of a long-term contract

A second alternative account for public utility regulation is the need for some form of administration of the long-term relationship between consumers and producers of public utility services.[64] On this view, regulation is not principally premised on the need for a regulator to determine efficient prices, but rather on the need for a body to administer, or govern, the terms of trade over a long-term contractual relationship between a public utility firm and its customers, in circumstances where there is uncertainty, and the relationship is complex and multi-dimensional.[65] In this context, the need for regulation arises because it is impossible to determine an optimal or complete contract at the outset. This line of reasoning is based on the idea that the contracting problem associated with public utility industries is a variant of a more general problem associated with long-term contracting in the context of uncertainty where parties incur relationship-specific investments (more specifically, where parties make durable and immobile investments).[66]

This characterisation of the essential problem that economic regulation is designed to address as one involving long-term contracting, has received increased attention in recent years and is being promoted by some as a useful way of conceiving of economic regulation as it operates in practice.[67] In this respect, the formal regulatory revenue determination and rate making process might be viewed as a form of 'dispute resolution'[68] of which there are other more informal alternatives. One of these alternatives, discussed in Chapter 3, is the negotiated settlements process used in North America, which involves the settling of rate cases by agreement between the public utility company and its customers and other stakeholders, typically without the involvement of a regulator, although any agreement

[63] See Joskow and Noll (1981:39); Peltzman (1989:58); Viscusi, Harrington and Vernon (2005:393).
[64] See, in particular, Goldberg (1976:431).
[65] Williamson (1976:103), for example, argues that rate of return regulation can be seen as 'a highly incomplete form of contracting in which the prospects for windfall gains and losses are strictly limited and, in principle, and sometimes in fact, adaptations to changing circumstances are introduced in a low-cost, nonacrimonious way'.
[66] See Goldberg (1976); Williamson (1976) and Gómez-Ibáñez (2003:9). See also older conceptions of contract management by Chadwick (1859), as discussed in Crain and Ekelund (1976:150).
[67] See Gómez-Ibáñez (2003) and Biggar (2009).
[68] Littlechild (2012b:174).

reached is subsequently submitted to the regulator for approval.[69] In this respect, some have argued that, by focusing on the traditional formal processes of revenue determination and rate making, economists have 'unduly neglected' alternative processes for settling such cases in some parts of the world,[70] and that, in fact, 'settlements between utility and consumer representatives are a major part of modern regulation'.[71]

This alternative characterisation of regulation has important implications for the role of a regulator, which changes from one in which it is tasked with representing the consumer interest and making final decisions on revenue determinations, to one in which it facilitates and enables well-informed participants to reach agreements that are mutually beneficial.[72] In effect, the regulator's focus is the process by which negotiations are conducted, rather than the outcomes of that negotiation process.[73]

2.2.3 Other potential explanations for the existence of public utility regulation

A more general explanation for public utility regulation relates to the importance of public utility industries to both an economy and to society. In this context, economic regulation is argued to reflect a political recognition that the pricing and allocation of such essential services are 'too important' to be left to market processes.[74] As Alfred Kahn observed, public utility industries have a 'public character' that is uniquely connected to the process of economic growth,[75] and the efficient provision of public utility services is likely to benefit a number of firms in other sectors in an economy. While this argument alone arguably cannot explain regulation of the public utilities – a range of non-utility activities also have economic and social significance – it usefully highlights the close connection between the public utility industries and the social, economic and political sphere.

A separate argument is that the conduct of regulation reflects the institutional interests of regulatory agencies themselves.[76] According to this reasoning, the continuing existence of regulation of specific public utility services can, in part, be explained as 'self-interest' on behalf of the regulatory agency – to 'stay in business' and maintain or expand its powers and jurisdiction. While this explanation cannot account for why regulation occurs in the first place, it does potentially have some explanatory power when it comes to considering why regulation is not necessarily withdrawn from those activities

[69] See Wang (2004:141); Doucet and Littlechild (2006:266).
[70] Doucet and Littlechild (2009:4633) go so far as to suggest that the traditional model for rate-setting set out in legal and economics textbooks is no longer the norm, but rather a fallback position.
[71] Littlechild (2009a:108).
[72] Doucet and Littlechild (2009:4643), for example, find that the use of negotiated settlements in Canada by the National Energy Board reflects this change: 'The prime role of the Board is no longer to impose its own view of the public interest. It is to enable well-informed market participants with a demonstrable interest to negotiate satisfactorily on something like equal terms with the oil and gas pipelines.' See also Littlechild (2012b:174) on the proactive role played by staff at the Federal Energy Regulatory Commission (FERC) in seeking to facilitate agreement between parties.
[73] See Doucet and Littlechild (2009:4640).
[74] See Scherer (1980:482).
[75] Kahn (1971:193).
[76] Generally, on this topic see Niskanen (1971).

where competition has developed, and the form that regulation has historically taken in different areas of the public utilities. In particular, it may suggest why the scope of activities pursued by many regulatory agencies in the public utility industries has tended to increase, rather than contract, over the past decades.[77]

Perhaps the most controversial alternative explanation for why public utility regulation exists, and takes the form that it does, is that economic regulation is, at least in part, a response to distributional issues, including issues relating to fairness and the protection of various groups in society. This view rests on an underlying assumption: that public utility services are services that should be provided to all citizens of a society and on a broadly equivalent basis.[78] On this line of reasoning, one of the functions of regulation is to ensure wide coverage and affordable access to public utility services. This rationale would explain why we see certain consumers cross-subsidising other consumers for public utility services. However, implementation of economic regulation premised on a prescription of 'fairness' is controversial, for at least two reasons. Firstly, economists arguably do not have a clear set of 'distributive weights' to allow issues relating to distribution to be dealt with in a systematic and non-arbitrary way in regulatory decision making.[79] Second, it is not obvious how considerations of fairness are, in practice, balanced against issues relating to economic efficiency in the context of the public utilities. For example, as discussed in Chapter 4, efficient forms of price discrimination to recover fixed costs typically involve applying higher mark-ups to customers with relatively low price elasticities. In some cases, this may result in a disproportionate burden being placed on those sectors of society whose demand is inelastic precisely because they have no real alternatives (i.e. working people who would pay higher transport fares).[80] This presents a potential and very real trade-off for a regulator between economic efficiency and equity issues in approving or disallowing particular pricing structures.[81] For these reasons, among others, economists have traditionally argued that public utility regulation should focus solely on matters of efficiency, and that to the extent to which issues relating to inequality or distributive justice arise in the public utility industries these are best dealt with through the taxation system or other redistributive policies.[82] However many, including

[77] Williamson (1976:75) discusses this tendency more generally. See also the discussion in Bernstein (1955:40) on the 'natural tendency' of the first federal regulator in the USA (the Interstate Commerce Commission, ICC) to seek to extend its powers. In its 2007 review of UK economic regulators, the House of Lords noted that it had received a lot of evidence highlighting that regulators' roles had kept expanding, and that this was taking them away from their eventual demise. See, HoL (2007:para 7.38). Shleifer (2011) argues more generally that the ubiquity of regulation in American and European societies may be explained by the failure of courts to resolve disputes cheaply, predictably and impartially.

[78] See the discussion in Helm and Yarrow (1988:iv) The premise is one that is adopted by some regulators. For example, the UK telecommunications regulator has described telecommunications services as so fundamental 'that all people, whoever or wherever they are, must have access to a certain basic level of telecommunications facilities and services if they are to participate fully in modern society'. Oftel (1997).

[79] See Schmalensee (1979:21). Contrast, Baumol (1986), who suggests that fairness is tractable to economic analysis, including in areas such as monopoly pricing and peak or congestion pricing.

[80] See Helm and Yarrow (1998:iv) and Baumol (1986:4).

[81] Zajac (1978:47) usefully describes the 'policy maker's dilemma' in setting efficient pricing structures.

[82] Schmalensee (1979:20) articulates the position more fully that regulators should not 'have to decide conflicts between efficiency and other goals'. See also Kahn (1971:68).

economists, would concede that, in practice, regulators, politicians and the courts do consider issues of fairness and distributive equity in applying regulatory policy.[83]

2.3 IMPLICATIONS OF THE DIFFERENT RATIONALES FOR REGULATION

This chapter has considered various accounts for why economic regulation might exist in the public utility industries, ranging from normative rationales based on the need to achieve efficiency, to accounts that focus on the interaction of different interest groups in society and how each group's interests may give rise to a demand for regulation and influence its form. Each of these accounts can potentially explain at least some aspects of regulation as it is applied and observed in practice. Yet, no single account seems to explain fully why regulation exists in the form that it does in many of the public utility industries.

While it is not possible to pinpoint a single unified and comprehensive account for public utility regulation, the recognition of different rationales and purposes is, itself, potentially illuminating. It may, for example, help to explain why it is that we observe multiple objectives in the remits of regulators of some public utility industries, rather than a single objective – such as to improve economic efficiency as the traditional normative theory of natural monopoly might suggest. It also helps explain some of the apparent tensions that exist between regulatory precepts of efficiency and ubiquitous regulatory practices such as cross-subsidisation. A recognition that the regulation of the public utilities is shaped by a range of concerns, interests and policy objectives is not, however, an endorsement of such a situation. As we will see in later chapters, a lack of clarity about the purposes of economic regulation can create challenges for regulators in understanding what they should be doing and, to paraphrase Alfred Kahn's quote at the start of this chapter, what action is 'really necessary'.

[83] See Berg and Tschirhart (1988:324), and Baumol (1986) for a discussion of issues of fairness. In practice, a concern for 'fairness' seems to be widely acknowledged by regulators in many parts of the world. The England and Wales water regulator has noted: 'Water customers … need to know that the bills they pay are fair and legitimate' Ofwat (2011a:2). Similarly, State Public Utility Commissions in the USA often refer to their remit as involving ensuring that 'regulated utilities offer their services to the public at a fair price' (Alabama Public Utilities Commission (2013:1)) and that citizens receive 'adequate, safe, and reliable public utility services at a fair price' (Public Utilities Commission of Ohio (2013:1)).