Performance Incentives for Network Rail: A Perspective from Behavioural Economics

Nick Chater

Executive summary

Conventional principal-agent models in economics understand the problem of regulation as that of a principal designing a system of rules and incentives. The agent is expected to choose its actions in order to maximise its expected utility, given those incentives, and within the constraints of the specified rules. The principal's task is therefore to design a mechanism to align the agent's incentives with the principal's own incentives.

From a behavioural point of view, this framework needs to be supplemented in a variety of ways:

1. ***Dealing with two types of risk.*** Individuals and organisations are often excessively averse to manageable risks that are inevitable in innovation and change. Yet they are also often insufficiently sensitive to hazards: low probability, and extremely negative, events, of which they may have no direct prior experience. Thus the agent may not behave as expected by the principal, unless these risk misperceptions are accounted for.

2. ***Trust and adversarial relationships.*** Where two or more individuals or organisations need to cooperate and coordinate their actions to achieve a common goal, the nature of the relationship between the parties is crucial. Particularly important is their ability to make ongoing, flexible, ad hoc agreements as their working relationship develops. One potentially crucial barrier concerns information asymmetry: if one party knows more than another about some issue (e.g., the state of some aspect of the rail network), then agreeing responsibility will be fraught with difficulty issues of trust. More broadly, it may be useful to establish contracts which focus on the nature of the relationship between the parties, including roles and responsibilities, rather than outcomes (e.g., cash transfers based on performance), which may lead to the parties concerned to adopt an adversarial, rather than cooperative, mind-set.

3. ***The logic of consequences and logic of appropriateness.*** There appear to be two very different frameworks in terms of which people evaluate their own and other people's behaviour. According to the logic of consequences, behaviour is judged by its results; often though, people do, and should, focus on a logic of appropriateness: the question is "what should a person in my role do, in a situation like this?" So, for example, issues of honesty, safety compliance, following professional norms and standards, acting as a good citizen to achieve common goals, and so on, are matters of appropriateness rather than consequence (although, in the long term, departures from such norms and standards will, of course, have very negative consequences).

4. ***Enhancing coordination and cooperation.*** Contracts, whether formal or informal, are fleshed out and implemented through a process of continuous renegotiation, and are often governed not merely by attempts to generate good consequences, but, as noted above, by requirements about what counts as "appropriate behaviour" (e.g., following professional

norms and standards; honest reporting; coordinating and cooperating to achieve goals whether or not incentivised by the contract; and general good citizenship). Traditional regulation by incentives (thus modifying consequences) tends to downplay questions of appropriateness; mechanisms to encourage a sense of common purpose, professionalism and integrity may be of considerable importance.

5. ***A richer model of motivation.*** Human motivation is complex, and can sometimes be undermined, rather than enhanced, by financial incentives. Moreover, people can be motivated by a wide variety of non-financial factors, including the intrinsic pleasure of completing a task successfully, seeing progress with respect to one's own previous performance, contributing to the common good, and gaining a reputation for "good citizenship."

Recommendations for further consideration and analysis

The wide range of behavioural factors surveyed in the report highlight the importance of encouraging a positive "social contract" between parties who are jointly working to improve the railway, encourage a sense of common purpose and teamwork, rather than being subject to incentives which encourage an adversarial culture. There should be considerable emphasis on encouraging appropriate norms of behaviour, rather than merely focusing on outcomes. And, as part of this 'social contract,' it is likely to be useful to agree broad measures of performance, with the potential for comparison with both one's own past performance and performance of other parties. Most important, perhaps, is creating a culture of public service, integrity, and cooperativeness in working to common objectives. Rather than being perceived as a burden, rigorous evaluation and scrutiny from the regulator would ideally be perceived as a helpful mechanism to achieve these common goals.

The regulation of a highly complex, politically-sensitive, publicly owned organisation such as Network Rail presents a number of challenges. It is, of course, vital that the regulator set out a framework of rules and incentives that will lead to a cost-effective and coordinated response from the many elements of Network Rail itself. It is also important that wider stakeholders, including the media, politicians and the general public can see that the regulator is providing tough and rigorous oversight of the industry. At the same time, in the context of a publicly owned body such as Network Rail, traditional financial penalties for unsatisfactory performance can be viewed as problematic, because they can be viewed as merely shifting public sector money from one organisation to another. It is therefore timely to consider a broader range of mechanisms that may be important in regulation, especially taking account of behavioural factors that may be crucial in promoting motivation, cooperation and coordination, adherence to safety and professional standards, and so on.

This review paper aims to survey relevant recent research in behavioural economics and related fields, and point to some potential links to regulation in the rail industry, and with respect to Network Rail in particular. The objective is to widen the range of considerations and options available, which may feed into ORR thinking in the development of specific regulatory mechanisms.

In the absence of behavioural considerations, the effect of performance incentives on managers and firms has traditionally been analysed using the tools of standard microeconomics, such as principal-agent models. In this framework, the manager is assumed to optimise a well-defined performance measure, often assumed to be personal salary; and the challenge of the principal (the organisation) is to align these personal incentives with the objectives of the organisation. From a behavioural economics perspective, this framework may need to be enriched and modified in the light of a number of factors.

The aim of the review is therefore to bring together behavioural and consider how they may bear on building a behaviourally robust performance management system. Crucially, in order to lead to practically useful recommendations, behavioural factors need to enhance, rather than replace, traditional economic analysis; and to be linked with a broader understanding of organizational norms and culture.

1. Dealing with two types of risk

In economics, the term 'risk' covers any aspect of the future that is not known.[1] Yet, from a psychological point of view, risk is complex. Slovic (1987) famously notes two key dimensions concerning the psychological perception of risk: concerning "unknowability" and "dread." Thus,

[1] Strictly, *risk* is sometimes reserved for cases where lack of knowledge can be quantified using a probabilistic model---e.g., in gambling, or in finance theory, where risk can be quantified. Cases in which no credible model is available, e.g., concerning the probability of financial crisis, a catastrophic failure of the grid, or wholesale rail renationalization, are then labelled as involving *uncertainty* (Knight, 1921). In real business contexts, credible probabilistic models are rarely available, and the distinction is of little practical use---though the dimension of "unknowability" is psychologically important (Slovic, 1987), as noted in the main text. We will therefore use 'risk' throughout.

for example, public fear of nuclear power is amplified by the unknown level of risk; and the "dread" of the imagined consequences of a major accident.

For this, and other, reasons, hazards where an outcome is rare and catastrophic (e.g., accident, prosecution, or financial collapse) are treated very differently from factors which merely have some inevitable degree of variability (e.g., date of project completion, cost, delays due to weather, and so on). We will call these risks of *hazard* and *imprecision*, respectively.

Risks of imprecision are inevitable. In the context of finance, indeed, risks trade-off against returns; so that a portfolio of risks is essential to achieving good overall fund performance. In non-financial businesses too, risks-of-imprecision will tend to trade off against expected return; and many such risks can be diversified across an entire organization. But the incentives of individuals may often encourage too little risk-taking---unexpectedly good performance is often rewarded (whether financially or informally) much less than unexpectedly poor performance is criticized. Indeed, such "loss aversion" is widespread in decision making—people tend to shy away from individual risks, because "losses loom larger than gains." The desire to "nail down" projects and contracts precisely, or to go on with 'business-as-usual' rather than experiment with an innovative way of working or a new technology, is likely, in aggregate, to lead to poorer overall organisational performance. It is therefore worth thinking about how to counteract excessive aversion to risks of imprecision--- for example, explicitly evaluating level-of-innovation might be one approach.

Hazard risks, by contrast, concern low probability, but very negative, events which the industry seeks to minimise as far as possible: serious accidents, systemic technological failures, or financial collapse. Here, the danger is the very rarity of the events concerned. People typically judge the probability of events, at least in part, by drawing from their own past experience, or the past experiences of those around them.

Laboratory studies of decision-making where people must estimate probabilities from experience (the so-called decision by experience paradigm, Hertwig, Barron, Weber & Erev, 2004; Ungemach, Chater & Stewart, 2009) have consistently shown that people tend to behave as if they severely underweight the probability of rare negative events. One reason for this is that such events are not even considered as possible at all---indeed, they may not ever have been present in a person's past experience, or only in the remote past (and people may then judge that "it couldn't happen now"). Thus, in the context of principal-agent interactions, there is the danger that incentives are set up which do not sufficiently deter behaviours that may lead to bad outcomes (e.g., underbidding on a franchise and subsequently being unable to deliver the contract). Here, the problem may be that the probability of rare events is underestimated both by the principal and the agent. Even where the principal correctly estimates the probability of a bad outcome, and sets up incentives which would properly incentivize a fully-informed agent, there is still the danger that the agent does not align with those incentives, due to underestimating the 'hazard' risks. To take an example from a different industry, a principal regulating safety procedures in the oil industry might hypothesize that no more than light touch monitoring and regulation is required to regulate safety at work, because accidents have such disastrous consequences for all involved. Experience is quite the opposite: without close monitoring, violation of safety-critical procedures can rapidly spread throughout a working culture. Because accidents are rare, long periods of poor safety culture and become established, and

reinforced by the lack of negative consequences, until a disastrous incident occurs. Here, lessons from research and practice on safety, while not the focus here, provide potentially important lessons.

This is the common pattern in "organizational disasters" where systemic multipronged safety failures have catastrophic outcomes (e.g., BP's Deep Water Horizon disaster in 2010, Reader & O'Connor, 2014). In that disaster, at least eight separate safety failures were involved; the internal procedures for guarding against such failures have themselves been eroded, particularly in the light of operational pressures. When regulating against the possibility of hazards, psychological biases can be circumvented by direct and independent monitoring of conformity with safety procedures (with severe penalties for violations); and a professional and reporting separation between those tasked with safety monitoring and implementation (in the airline industry, the ability of an engineer to ground a plane, irrespective of commercial considerations, has been crucial to raising safety standards).

But hazard risks go well beyond safety: failure to deliver a project successfully, or with catastrophic overspend, or financial collapse of some part of the business, are among the hazard risks to be considered. A regulatory challenge is to consider whether sufficient professional and line-management separation between risk assessment and delivery is in place to reduce such hazard risks. Moreover, work in the oil industry (Reader, Mearns, Lopes & Kuha, 2017) suggests that promoting a 'safety culture' requires employees to feel valued and supported by their employers---the same may be true regarding the focus on hazard-risks, whether connected to safety or not, of the employees of Network Rail and its partners.

2. Trust and adversarial relationships

Any incentive structure proposed by the principal to govern and optimise the behaviour of the agent will constitute a formal or informal *contract*. From the point of view of conventional economic theory, the agent's response to the contract will be to behave in a way that has the best expected consequences for the agent. As we have noted, then the challenge for the principal, in designing the contract, is simply to align the incentives of the agent with those of the principal as far as possible.

One reason that designing such contracts is difficult, according to a conventional account, is information asymmetry: for example, the agent may be much better able to assess its own performance compared with the principal. For example, the business unit concerned with maintaining the quality of the track may be far better able to determine its responsibility for the degree to which trains are late running, than the rail operating companies. An incentive structure requiring payments proportional to responsibility is, of course, likely to distort the reporting of that performance. The suspicion that underreporting may occur by the counterparty---the train operating company---is likely to create a low sense of trust and an adversarial relationship between those responsible for track maintenance and train operation. While the difficulties of information symmetry are within the scope of traditional economic models concerned with "mechanism design," the resulting adversarial relationship between the parties may have considerable behavioural spill-over. That is, once pitched into an confrontational relationship, where trust is low, agents may struggle to coordinate and cooperate successfully, even in aspects of the business where such cooperation is in the interests of both parties.

Adversarial relationships arising especially in situations in which agents consider success is defined relative to each other, rather than in absolute terms. Consider two agents A and B, who have the opportunity to accept a deal which benefits A by 5 units and B by 1 unit. If A and B have a positive and trusting relationship, they may both happily accept (B will be further encouraged to accept the deal because of its large benefits to A; and B will anticipate that A will likewise accept any future deals were B is the main beneficiary). By contrast, if A and B have an adversarial relationship, the large benefits to a A will deter B from accepting, even though B benefits; and B will not be confident that A would accept any future deal for which B would be the main beneficiary). ***Adversarial relationships can, therefore, lead to bad outcomes for both parties; and hence avoiding incentives and rules that encourage such relationships is of considerable importance.***

Moreover, informational limitations may apply both to the principal and the agent: neither may have, for example, accurate measures of "efficiency." Often, indeed, many of the most important measures of success are the most difficult to quantify. It is of considerable importance that difficult to measure but crucially important factors are not neglected. Moreover, there is likely to be the greatest latitude for dispute between principal and agent concerning such measures. It is therefore advisable, where possible, to have independent assessment of difficult-to-quantify measures. In some context, this assessment can usefully be provided by peers or fellow stakeholders. For example, a reputation for "good citizenship" can be measured by brief qualitative evaluations from, for example, short web-based questionnaires given to those stakeholders. It is worth stressing that people are intrinsically motivated to obtain the esteem of their peers: gathering and publishing such feedback may considerably reshape behaviour, even if not tied to any financial incentive (for example, I know from personal communications, that this type of mechanism has been used successfully to encourage cooperation between senior managers across geographies in a global bank---managers are apparently very keen to be perceived by their peers as helpful and constructive—at least once this behaviour is measured).

3. The logic of consequences and logic of appropriateness

Traditional principal-agent models focus on the problem on incentive systems that encourage agents to choose the action with the best consequences. But a richer conception of human behaviour suggests that equally, or perhaps more, important is the effect of incentives on orienting agent to focus on consequences *at all*.

March and Olsen (2004) importantly distinguish two very different types of motivation for human behaviour. Conventional economic theory focuses on what they term the logic of consequences: actions are evaluated by their effects. According to the logic of consequences, each agent asks: "what should I do, in the light of the opportunities and incentives that I face, in order to yield effects that align with my goals?" This viewpoint is formalised in the assumption that agent choose their actions to maximise their expected utility (this viewpoint does not, though, require that agents are purely self-regarding; agents' goals might, for example, be concerned with the well-being of others).

By contrast, a great deal of human behaviour has a very different origin. According to the logic of appropriateness, each agent asks "what is an agent like me supposed to do in a situation like this?" The logic of appropriateness is to the fore when we consider norms, standards, roles and

responsibilities. Thus, in a trial, what the defence is "supposed to do" is to present the strongest possible case against conviction; the prosecution is, by contrast, supposed to prevent the strongest possible case for conviction; the judges are supposed to ensure fair play and to sum up impartially, and so on. It is expressly not appropriate for each party independently to choose what they say in order to achieve what they happen to believe is the "best consequence" (i.e., the defence should not weaken their rhetoric, if they happen to doubt their client's case). The question of determining guilt is, after all, the role of the jury.

For example, in the present context, the logic of appropriateness might dictate that a safety inspector check each portion of track at a set frequency and to a set standard; the logic of consequences might suggest particular, problematic, portions of track should be prioritised. But the rules that govern such safety inspections will, and arguably should, determine the inspector's behaviour, at least until the point at which these rules are modified. More broadly, in any organisation, people have to learn, or infer, the "ways things we do things round here" and "what someone in my role is supposed to do" rather than adopting a general consequence-based approach deciding how to act.

Indeed, as in the context of a court, the well-functioning of many processes within, and between, organizations depends on people and companies being guided by the logic of appropriateness: in essence, following the rules, standards, and guidelines appropriate for their role. Thus, behaviour is guided by following the rules of the process rather than being guided directly by the outcome. For example, roles concerned with safety, audit, or reporting in general typically involve following professional and legal standards. But more broadly, to the extent that the principal and agent (or a group of agents) are governed by a "contract" then the required behaviour is primarily to fulfil the contract to the required standards (e.g., not the barely fulfil merely the letter of the contract, but not its spirit; or to fulfil the contract only where verification is likely, and so on).

Many aspects of the behaviour of both individuals and organizations concern the balance between the role and scope of the logic of consequences and the logic of appropriateness. Performance-based incentive schemes focus on consequences---how the consequence is achieved is secondary. Of course, the assumption is, reasonably, that targets should be achieved within regulatory and legal standards. But this is insufficient because (a) as we have noted, the fulfilment of those standards can become partial and perfunctory, which is likely to be undesirable; (b) however detailed the "contracts" expressing those standards, they will inevitably be open-ended. There will be many cases where judgement is required about what "is appropriate"---and performance-based incentives will tend to corrode the application of such standards. For example, in many areas of business, salespeople are incentivised by volume of sales; and while they may be required to sell only to customers who have a genuine need for the product (e.g., notoriously when selling PPI), such requirements can easily become marginal or ignored entirely, in the face of the incentive to sell. Even when money is not involved, a performance-based target can encourage people and organizations to operate as if they are playing a 'game' with a well-defined objective, and that the 'rules' of the game need only be respected in a perfunctory way, or even subverted entirely.

In this light, a key question for a regulator is to degree to which standards should focus on 'consequences (targets, levels of output/performance etc) or 'appropriateness' (incentivising types of 'good,' professional, cooperative, behaviour). The degree to which appropriateness can be

measured is limited: the judgement of peers and stakeholders, rather than objective data, is likely to be most appropriate. Such measures of the professionalism, good citizenship, cooperativeness that pick up the agent's or organization's reputation for behaving appropriately will be crude and imprecise (for example, people's judgements of each other and of businesses often reflect little more than a few dimensions. But they have the virtue of being difficult to 'game'---the best way to build a reputation for, e.g., honest reporting, is almost certainly to engage in honest reporting. (And any hint that dishonest reporting is being covered up creates great reputational risk).

4. Enhancing coordination and cooperation.

Contracts, regulations, and professional standards are inevitably open-ended: there will always be unforeseen circumstances in which the appropriate way to act is not well-defined. Indeed, a great deal of legal scholarship and controversy arises from such open-endedness. In the context of laws governing the individual, controversy concerns what, for example, precisely follows from a 'right to family life,' what amounts 'informed consent' for sharing of data (particularly in the age of pre-populated tick-boxes in terms and conditions), and so on. Such open-endedness will pervade formal regulations and informal norms alike in the rail industry, as in any other: and such open-endedness allows for differences of opinion, and hence potential dispute, concerning what the counterparties are committed to. Hart (1995) suggests that the incomplete nature of contracts, and the continual possibility of renegotiation, is a central aspect of economic life---and will, therefore, arise in any contracts set up to govern principal-agent interactions, which will be potentially open to dispute and renegotiation.

There are directly contrasting approaches to dealing with such open-endedness. One approach is to attempt to pin down the rules as precisely as possible. A second approach, sometimes known as 'relational contracting' (in contrast to 'transactional contracting') aims instead to establish general roles and responsibilities, agreed objectives, mechanisms of interaction and dispute resolution, and to build a relation of trust between the contracting parties[2] (e.g., Frydlinger, Cummins, Vitasek & Bergman, 2016; MacNeil, 1968).

This type of approach typically focusses on encouraging a non-adversarial culture, sharing risks, and finding mechanisms jointly to solve problems as they arise. One well-known example of the effectiveness of this approach was the contract used by BAA in the construction of Heathrow Terminal Five, completed on time and within budget (Carter & Mukhtar, 2008); and a balanced scorecard methodology was an important part of the performance management approach underpinning the agreement (Basu, Little & Millard, 2009).

A rich literature in behavioural game theory has explored some of the factors that determine whether agents are able to work together successfully. Traditional principal agent models typically assume that the most crucial factor is that the parties have common, or at least well-aligned, goals. Equally, and often more important, though, is common knowledge: having a shared understanding of the

[2] To quote Frydlinger et al, a relational contract is "A legally enforceable written contract establishing a commercial partnership within a flexible contractual framework based on social norms and jointly defined objectives, prioritizing a relationship with continuous alignment of interests before the commercial transactions."

nature of the challenge to be addressed, the roles and responsibilities of each party, mechanisms by which disputes should be resolved, and so on.

Critical to successful coordination is the ability to "team reason" (Bacharach, 2006; Sugden, 2003)---that is, to ask "what would we agree is the right thing to do, in a situation like this?" Successful team reasoning involves spontaneously, and without communication, coming up with a common course of action, including who responsible for what, who should take which actions and bear which costs. Where team-reasoning is possible, explicit negotiation will be straightforward--- indeed, in some cases, explicit negotiation may not even be required (this can be particularly important in time-critical decisions, for example, in emergencies). By contrast, if the parties independently form two very different views about what "we" should all do, then spontaneous coordination will be poor, and negotiation fractious. Common objectives and common knowledge are therefore crucial for successful team reasoning.

Consider, for example, the following abstract problem. Two players see a number of sums of money sitting on a table top; the players sit at opposite ends of the table. They can independently select one or more of the sums of money. If they both select any of the sums of money, then neither player gets in any payoff (we might think of this as analogous to beginning a mutually damaging conflict over customers, land, or any other resource). If their choices do not overlap, then both players receive the sums they have chosen. Thus, the players will succeed if they are able independently to decide how "they" should split the sums of money between them---if they cannot agree, then either money will be "left on the table" or, more likely, their choices will overlap and neither player will receive anything (this set up is roughly that studied by Isoni, Poulsen, Sugden & Tsutsui, 2013, 2014). Notice, crucially, that mere goodwill is not sufficient to solve coordination problems like these: both parties might be willing to share the sums of money equally, but they crucially have independently to generate the same specific allocation of sums to players. In some contexts, this is straightforward: if half the money positioned close to Player A's end of the table, and half the money is positioned close to Player B's end of the table, then both players may 'agree' (though without communication) to select the sums of money nearest them. But notice how common knowledge is crucial: this tacit agreement cannot be reached if the players don't know their own location at the table, know the other player's location, know that the other player knows both of these things, and so on. Similarly, if the locations of the sums of money are not common knowledge, then coordination will be impossible. Moreover, common knowledge of individual preferences or abilities, past experience, of conventions may be crucial. Suppose that, instead of sums of money, two parties have to divide two types of good (e.g., apples and oranges); then common knowledge that A prefers apples and B prefers oranges would radically simplify the problem of spontaneously alighting on the same "agreement" about sharing the goods. Indeed, without such common knowledge, the task may be impossible.

In the context of principal-agent interactions, especially with multiple agents, lack of common knowledge may be at least as much of a barrier to harmonious interaction as are conflicting goals (e.g., that neither agent wishes to accept responsibility for, or pay for, a failure in the system).

Moreover, the ability to team reason successfully requires substantial effort and attentional resources---each party has to attempt to understand the challenges and constraints of the others, and

to imagine what reasonable agreement might be reached (Misyak & Chater, 2014; Misyak, Melkonyan, Zeitoun & Chater, 2014). The ability to focus attention on the other's perspective and to create a common understanding of what is jointly "reasonable" has been termed 'social mindfulness' (Van Doesum, Van Lange & Van Lange, 2013; Van Lange & Van Doesum, 2015). It is important to distinguish social mindfulness from other-regarding preferences (e.g., altruism). In order to work successfully together, parties need to seek to *understand* each other, and the common problem they face, not mere to sympathize with other. Creating forums in which information can be shared and a rich basis of common knowledge can be established may be of key importance.

5. A richer model of motivation.

One of the main purposes of principal-agent models is to capture how the motivation of the agent will lead its choices, and hence how the principal should adjust incentives to shift the motivation of the agent so that its behaviour aligns with the goals of the principal. Both principal and agent are assumed to be rational utility maximisers. This type of framework is useful for considering some ways in which incentives can lead to 'perverse' outcomes. For example, performance incentives with "ratchets" or may encourage initially low performance, which can then be improved upon. Equally, and potentially importantly, comparison and competition between agents may in some circumstances lead to a breakdown of collaboration, and even behaviours by which each agents aims not merely to improve their own performance, but to damage the performance of other agents. For example, in sales, individual incentives can lead to turf wars, stealing of customers, and so on between sales people in a way that is against the interests of the organisation at large. Similarly, tying incentive to a particular measure (e.g., the percentage of A&E patients who are seen within four hours in NHS hospitals; or the proportion of pupils with five A-C grade GCSEs) can notoriously lead to unwanted behaviours, such as focussing effort at boundaries, shifting attention away from non-measured outcomes (including those contributing to the common good), and potentially undermining professional norms and procedures which may be in tension with hitting the measured target (Gray, Micheli & Pavlov, 2014).

In addition to these considerations, it is crucial to have a richer and more realistic conception of motivation, which allows for the possibility that the incentives provided by the principal may not moderate the agent's behaviour to pursue those incentives at all (for a review, see Bowles, 2008). In particular, given that external incentives can crucially undermine intrinsic motivation, they must be applied with great care.

To focus on a particularly well-known behavioural study, Gneezy and Rustichini (2000) found that the introduction of fines for late pickup at an Israeli nursery school had the perverse effect of increasing the degree of parental lateness. They argued that many parents' interpretation of the new incentive was that "a fine is a price." That is, rather than feeling obligated to fulfil their part of an implicit bargain with the nursery school to pick up their children on time, these parents viewed the fine as providing a legitimate way in which they could, where required, extend the nursery school day by providing extra payment. Moreover, when the policy was reversed, the increased levels of lateness continued: it appeared that once lateness had been legitimised, a sense of obligation and social stigma for lateness could not easily be reintroduced. A contract-based viewpoint may be helpful to understand this behaviour. The school's interpretation of the implicit contract with parents

was that on-time pickup is mandatory, and the additional fine was presumably viewed as reinforcing the importance of this obligation; many parents, though, appeared to have interpreted the shift in policy as instituting a new type of market transaction: that the nursery would stay open, and presumably be happy to stay open, when suitably financially compensated by the parents. That is, they were interpreting the fine as a price that the nursery was willing to accept for late pickup. This transition can be viewed as shifting lateness from the domain of the logic of appropriateness (the nursery expects that, as a responsible parent, I should pick my child up on time) to the logic of consequences (my financial transaction with the nursery should be optimised in the same way as my other market transactions, so that I should "buy" additional nursery care when it is in my interest to do so). Gneezy and Rustichini notes that, with a transactional framework, the behaviour could be deterred with a suitably high fine; but then, of course, relations of trust between parents and nursery likely to decline rapidly. For example, if a parent is unavoidably late due to illness or unavoidable traffic congestion, then the demand for a large fine from the nursery is likely to lead to acrimony. Moreover, the very prospect of a large fine may pitch the parents and nursery in what appears to be an adversarial relationship, where the parent may perceive the nursery as benefiting from, and indeed wishing to encourage and exaggerate, their lateness.

A closely related pattern of "back-fire" behaviour is observed when previously voluntary acts of good citizenship are rewarded financially. For example, surveys and experiments on blood donations found that financial incentives are disliked by donors in New Zealand (Howden Chapman et al. 1996) and Sweden (Mellstrom & Johannesson, 2008), with especially strong negative effects for women. This negative impact appears to be sensitive to the precise details of the payments: for example, one Italian study finds that financial incentives are acceptable to blood donors if given in the form of a voucher rather than cash (Lacetera & Macis, 2010). One possibility is that this reduces the sense of the donation being treated as interchangeable with other forms of 'work.'

From the point of view of the principal, it is also important to be able to amplify, rather than undercut, intrinsic motivation. The literature on intrinsic motivation is vast. It is useful, though, to summarise some points, which are illustrated particularly well by people's choice of leisure activities---many of which will tend to be activities that people find intrinsically motivating, and for which there are few if any external incentives.

- People are often highly motivated by the purely intrinsic process of achieving a goal, creating a system or structure, or creating orderliness, irrespective of any wider purpose. So, for example, many computer games and hobbies involve setting and completing arbitrary goals (completing successive levels in a computer game; achieving a certain time or distance in running, cycling and rowing, etc.), or the creating order out of a measure of chaos (Farmville, crosswords, Sudoko, arts and crafts, and so on). It is important that any financial incentive does not undercut, or appear to devalue, the sense of inherent pride and achievement in "doing a good job."
- Prosocial motivations are commonplace. The sheer prevalence of volunteering illustrates the power of this motivation to do something either for the common good, or for the good of specific other people.
- These prosocial motivations are amplified when people feel part of a team. Indeed, the mere existence of the team can induce a very appealing sense of having a common purpose, even

when that common purpose is somewhat arbitrary, for example, in winning a football game in a local amateur league. The popularity of playing team sports is one illustration. Social psychology experiments have shown that even arbitrarily linking people together by, for example, giving half yellow T-shirts and half blue T-shirts, automatically amplifies cooperative behaviour within a team (for example, in the prisoner's dilemma game), although it can also reduce cooperation between members of opposite 'teams.'

- People are, moreover, typically motivated to improve on their own past performance. Again, leisure activities provide an elegant illustration of the power of such forces, where no outside incentives appear to be operative. In amateur sport, people continually attempt to beat their own personal bests, or achieve other targets (e.g., completing a marathon) relatively independently of the performance of others.
- Often, though, people are also concerned with how their performance compares to that of others. Indeed, as the economist Robert Frank and others have stressed, much activity and spending both at the level of individuals and companies may focus on so-called positional goods---that is, goods for which one's rank in relation to others is the primary objective. This is transparently the case in sporting competition, of course, where medals are awarded depending on relative performance; but arguably it may implicitly 'conspicuous consumption,' including the choice of expensive head offices, and so on (e.g., Frank, 1985, 2012). This focus on comparative rankings fits with the observation that league tables of all kinds can potentially be motivating for individuals and organisations. It is crucial, though, that the league table also allows evaluation of performance in absolute, rather purely relative terms: in this way, people and organisations are motivated to improve, and not encouraged to give up, where the general standard of their peers is improving, and where they may not be improving so rapidly. Indeed, it is important that across-the-board improvement in the sector is viewed as a positive development by all concerned, rather than a treadmill that is speeding up continuously. Incentives and reporting of performance needs to reflect this.
- Relatedly, one virtue of a basket of performance measures is that the morale of all parties may be maintained by finding that they are performing well on at least some measures. Of course, it is important that some overall summary of performance is also present, so that an appropriate balanced scorecard is achieved, rather than some individuals or organisations merely abandoning targets that they find difficult to meet. Nonetheless, flexibility for individuals and organisations in highlighting their strengths is likely to be important for good morale: indeed, people tend to react badly to feeling that there entire performance is, ultimately, distilled into a single number.
- Finally, as has been noted above, reputation matters. People are generally motivated to be seen as good and effective citizens within their personal lives and workplaces. The possibility of improving their reputation, and that of their organisation, will often be inherently motivating, although this will be strongly affected by the degree to which they feel that reputation is, at least partly, within their control. Such factors are likely to be more important than relatively small performance bonuses to individuals or teams---indeed, such bonuses may be seen as primarily of symbolic rather than monetary value.

Overall, then, it is crucial that the principal does not assume that the agent will be motivated purely by financial incentives defined by the principal. Indeed, such incentives may be secondary to

motivation coming from intrinsic or social sources, and may even interfere with those other sources of motivation in some cases.

Recommendations for further consideration and analysis

Behavioural factors concerning issues such as risk, trust, coordination and cooperation, norms and standards, and intrinsic motivation, suggest that a straightforward application of classical principal-agent models can have unexpected and potentially perverse consequences. Many of these factors may be at play in a single scenario. To illustrate, consider a striking laboratory experimental study by Falk and Kosfeld (2006), in which the agent selects how hard to 'work,' where the incentives are for the agent work as little as possible, whereas the principal wishes the agent to work as much as possible. Crucially, the principal can choose whether or not to impose a low minimal work level for the agent. When the principal imposed such a minimum, the agent tended to respond by working at or near that level, reporting a sense of lack of trust by the principal and a lack of autonomy. By contrast, when the principal allowed the agent choose any level of work, the agents worked substantially harder, on average, even though they now had the option of not working at all.

Even in this simple scenario, a range of complex factors is at work. For example, the principal may inappropriately be attempting to minimise *risk* or uncertainty by imposing a minimum level of work; indeed, in a real regulatory context, one can imagine the principal's incentives working in favour of imposing such minimum, on pain of being perceived as 'weak' by outside parties, even if the minimum is known to reduce the expected level of work from the agent. Moreover, the question of *trust* is crucial: by being trusted, the agent feels obliged to respond constructively---that is, the rate of *cooperation* is increased. An agent not responding in this way to the trust of the principal would be seen as violating social *norms*, and behaving inappropriately; and, finally, the result of these factors appears to be that the imposing a minimum performance level can undermine the *motivation* of the agent.

Despite these complexities, a good guideline for regulation in this industry, and elsewhere, is likely to be to create norms, standards, and incentives which are, as far as possible, perceived by all sides as fair; where the logic of appropriateness, and faithfulness to the spirit as well as the letter of any contract between the parties, is stressed as far as possible; and where a sense of common purpose for the public good is to the fore, and likely to be used as a guide for behaviour, so that perverse incentives are exploited as little as possible. Moreover, the collection of transparent data about performance, with the aim of raising the standards of the entire industry, but also allowing different parts of the industry to compare performance, for example across regions, is likely to be important. Tying the measurement of such performance to financial incentives needs to be treated carefully, to maintain the sense of working to a common purpose for which cooperation and coordination can be mutually beneficial, rather than being an adversarial relationship with other parties.

Ideally, whatever monitoring, evaluation, and incentive systems are developed, these should be perceived as agreed by, and treating fairly, different components of Network Rail, and also to be aligned with the concerns of external stakeholders. Rather than encouraging each part of the business to focus narrowly on its own financial targets, it is likely to be more productive to focus

on creating a public service ethos, with a culture centred on integrity, cooperativeness, and working towards common objectives.

The complexity of the range of behavioural effects that can play a role in determining the impact of regulation implies that, while prior literature such as that surveyed here can provide a useful starting point for the development of proposals, such policies will need to be openly discussed with the relevant parties, where possible carefully trialled, and modified over time in the light of experience. In time, it may be hoped that a rigorous process of regulation may be perceived by all elements of National Rail not as an additional burden, but as a valuable tool for helping to achieve the common objective of improving the UK rail network.

Issues for further consideration

Are there areas of network rail where high-risk/high-return innovation is discouraged by asymmetric incentives?

Are there low-probability but high-importance hazards (whether concerning safety, financial stability, or others) that may be under-prioritized (perhaps inadvertently) in the current environment?

Do incentives potentially create adversarial relationships between parties in the rail industry? How can this be minimized?

How far, and in which contexts, should Network Rail aim to shift from transactional to relational contracts?

What would be the impact of league tables on cooperation and coordination across different elements of Network, and external stake-holders.

How could ‘good citizenships’ be simply measured, e.g., using peer or independent assessment, both for individual managers or elements of Network Rail.

Are financial incentives undermining intrinsic motivation, and professional norms and standards?

How can the regulator help build sense of common purpose across the industry, and strengthen the commitment of all parties to public service?

References

Bacharach, M. (2006). *Beyond individual choice: teams and frames in game theory*. Princeton, NJ: Princeton University Press.

Basu, R., Little, C., & Millard, C. (2009). Case study: A fresh approach of the Balanced Scorecard in the Heathrow Terminal 5 project. *Measuring Business Excellence*, *13*(*4*), 22-33.

Bowles, S. (2008). Policies designed for self-interested citizens may undermine" the moral sentiments": Evidence from economic experiments. *Science*, *320*(5883), 1605-1609.

Carter, K., & Mukhtar, A. (2008). Partnering Heathrow Terminal 5. In K. Carter, A. Kaka, & S. Ogunlana (Eds.), *Proceedings of Joint International CIB Symposium,* Dubai. https://www.irbnet.de/daten/iconda/CIB17591.pdf

Falk, A., & Kosfeld, M. (2006). The hidden costs of control. *American Economic Review*, *96*(5), 1611-1630.

Frank, R. H. (1985). *Choosing the right pond: Human behavior and the quest for status*. Oxford University Press.

Frank, R. H. (2012). *The Darwin economy: Liberty, competition, and the common good*. Princeton University Press.

Frydlinger, D. Cummins, T., Vitasek, K. & Bergman, J. (2016). Unpacking relational contracting. White Paper: Haslam School of Business. http://www.vestedway.com/wp-content/uploads/2016/10/Unpacking-Relational-Contracting_v19.pdf

Gneezy, U., Meier, S., & Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *Journal of Economic Perspectives*, *25*(4), 191-210.

Gneezy, U., & Rustichini, A. (2000). Pay enough or don't pay at all. *The Quarterly Journal of Economics*, *115*(*3*), 791-810.

Gray, D., Micheli, P., & Pavlov, A. (2014). *Measurement madness: recognizing and avoiding the pitfalls of performance measurement*. John Wiley & Sons.

Hart, O. (1995). *Firms, Contracts, and Financial Structure*. Oxford: Oxford University Press.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*(*8*), 534-539.

Howden Chapman, P., Carter, J. and Woods, N., 1996: "Blood Money: Blood Donors' Attitudes to Changes in the New Zealand Blood Transfusion Service," *British Medical Journal*, *312*, 1131-32.

Isoni, A., Poulsen, A., Sugden, R., & Tsutsui, K. (2013). Focal points in tacit bargaining problems: Experimental evidence. *European Economic Review, 59*, 167-188.

Isoni, A., Poulsen, A., Sugden, R., & Tsutsui, K. (2014). Efficiency, equality, and labeling: An experimental investigation of focal points in explicit bargaining. *American Economic Review*, *104*(*10*), 3256-87.

Knight, F. H. (1921) *Risk, Uncertainty, and Profit*. Boston, MA: Houghton Mifflin Company.

Lacetera, N., & Macis, M. (2010). Do all material incentives for pro-social activities backfire? The response to cash and non-cash incentives for blood donations. *Journal of Economic Psychology*, *31*(4), 738-748.

Macneil, I. R., (1968). *Contracts: Instruments for Social Cooperation*. Hackensack, NJ: F. B. Rothman.

March, J. G., & Olsen, J. P. (2004). The logic of appropriateness. In *The Oxford Handbook of Political Science*. Oxford: Oxford University Press.

Mellstrom, C. and Johannesson, M., 2008: "Crowding Out in Blood Donation: Was Titmuss Right?," *Journal of the European Economic Association*, *6*(*4*), 845-63.

Misyak, J. B., & Chater, N. (2014). Virtual bargaining: a theory of social decision-making. *Philosophical Transactions of the Royal Society B*, *369*(*1655*), 20130487.

Misyak, J. B., Melkonyan, T., Zeitoun, H., & Chater, N. (2014). Unwritten rules: virtual bargaining underpins social interaction, culture, and society. *Trends in Cognitive Sciences*, *18*(*10*), 512-519.

Reader, T. W., Mearns, K., Lopes, C., & Kuha, J. (2017). Organizational support for the workforce and employee safety citizenship behaviors: A social exchange relationship. *Human Relations*, *70*(*3*), 362-385.

Reader, T. W., & O'Connor, P. (2014). The Deepwater Horizon explosion: non-technical skills, safety culture, and system complexity. *Journal of Risk Research*, *17*(*3*), 405-424.

Slovic, P. (1987). Perception of risk. *Science*, *236*(*4799*), 280-285.

Sugden, R. (2003). The logic of team reasoning. *Philosophical Explorations*, 6(3), 165-181.

Ungemach, C., Chater, N., & Stewart, N. (2009). Are probabilities overweighted or underweighted when rare outcomes are experienced (rarely)? *Psychological Science*, *20*(*4*), 473-479.

Van Doesum, N. J., Van Lange, D. A., & Van Lange, P. A. (2013). Social mindfulness: Skill and will to navigate the social world. *Journal of Personality and Social Psychology*, *105*(*1*), 86.

Van Lange, P. A., & Van Doesum, N. J. (2015). Social mindfulness and social hostility. *Current Opinion in Behavioral Sciences, 3*, 18-24.