The Problem of Business Evil*

John Thanassoulis †

February 14, 2025

Abstract

We explain why good people can be the cause of corporate scandals, such as *Oxycontin*, *WorldCom*, *Bernie Madoff* and *Societe Generale*. An ethical decision maker (other-regarding preferences) discovers privately that an action he thought was benign is harmful. This will become public knowledge at an uncertain future time. Stopping entails fines, carrying on means knowingly harming others, greater fines when the public signal arrives, but the possibility of escape via merger or death. I show that all decision makers, however ethical, have an *evil time horizon* – if knowledge the activity is harmful arrives after this time then the decision maker will select denial and knowingly choose to harm others – until caught. We show fines can make the problem worse, characterise alternative welfare enhancing policies, and demonstrate that ethical agents will lose a race to innovate.

Keywords: Misconduct, fraud, fines, finance, pharmaceuticals

*My thanks to audiences at the keynote address for the Hellenic Finance and Accounting Association conference 2024 and at Warwick Business School. All views and any errors are mine.

†University of Warwick, Warwick Business School, CEPR, UK Competition and Markets Authority. Email: john.thanassoulis@wbs.ac.uk

1 Introduction

It is not only bad people who do bad things. It is tempting to explain away every corporate scandal as an inevitable outcome of a key business executive being wicked. This paper explains why good people, even very good people, can knowingly choose to do bad things. A general trait of humans is that most of us are good. We espouse ethical values which include a preference for others not to suffer (Haidt (2007), Greene and Haidt (2002), Fehr and Schmidt (1999)). Nonetheless the capacity that each of us has to do very bad things is well known. It has been recognised in literature (e.g. "*Lord of the Flies*," by William Golding), in philosophy (e.g. Hobbes[1]), and is a core part of all major world religions (Christianity, Islam, Buddhism for example are all concerned that an over-emphasis of the self – a preference for one's own benefit – ultimately leads to suffering.[2]).

This paper explains why people with a strong sense of ethics can find themselves knowingly perpetuating behaviour they know is bad and harmful. Some examples of people who at the time, and sometimes still, were thought of as good people and yet who knowingly committed bad business acts include:

1. **The Sackler family**, continued to sell Oxycontin even though Purdue Pharma, their company, acquired evidence that their drug was highly addictive (similar in fact to the potency of heroin[3]) – Keefe (2021). Whilst Oxycontin remained on the market it contributed to an opioid crisis in the United States which ultimately saw nearly $\frac{1}{2}$ million deaths from overdoses between 1999 and 2019.[4] The original Sackler brothers were highly thought of doctors who funded the arts heavily including a Sackler wing at the Metropolitan Museum of Art in New York. Yet the family fought in court to avoid admitting that their drug was a problem and continued the denial and the sales for decades until the law caught up with them.

[1] Recall that in Leviathan, Hobbes (see Hobbes and Missner (2016)) worried that our natural state of nature was to be nasty and brutish.

[2] See references in Leary (2004)

[3] See Ohio Detox Centre for example.

[4] See Sackler family to pay $6bn for role in US opioid crisis, BBC, 3 March 2022.

2. **WorldCom Scandal**:[5] The head of accounting was Mr Buford Yates Jr. who reassigned opex as capex allowing WorldCom to defraud investors by maintaining fictionally high profits. The scheme was eventually exposed leading to huge losses amongst the investing public. Mr Yates makes clear that the personal cost to him of behaving unethically was substantial as he recounts "*there are no words to describe my shame and humiliation.*"

3. **Bernie Madoff** latterly ran two investment firms – a stock brokerage business and an asset management business. The asset management business began as a Ponzi scheme in which high payouts were initially funded by new joiners. But it became apparent to Mr Madoff that real investment results could not catch up with the fiction used to entice new entrants. Nonetheless Madoff continued until he was ultimately found out. Before this became public Mr Madoff was seen as a good man – he had a wife and two children,[6] he gave significant sums to charity through his multi-million dollar foundation[7], and at trial he declared his repetence for what he had done: "*I am so deeply sorry and ashamed [of my crimes] ... I cannot adequately express how sorry I am for what I have done.*"

4. **Jérôme Kerviel** and Societe Generale financial investments fraud. Jérôme tricked the risk control system so as to permit himself to place unhedged bets on equities. Initially his bets made a lot of money (so much that he felt he needed to make some deliberately losing bets to avoid drawing attention to himself).[8] However the bets then started losing money and eventually this became so substantial that Jérôme was discovered. He was sentenced in court to pay €4.9billion. Yet all of this was seemingly not motivated by greed: Forbes reports "*he apparently made no personal profit from his trades*".[9]

[5]WorldCom Figure Is Sentenced, The New York Times, Aug 10 2005.

[6]Though the elder son died by suicide blaming his father's deceit as the cause – Madoff Talks review: definitive life of an 'extraordinarily evil' man *The Guardian* 9 May 2021

[7]Standing Accused: A Pillar of Finance and Charity, *The New York Times*, December 12, 2008

[8]Jérôme Kerviel: History and Work With Derivatives, Investopedia, June 26 2022

[9]Pity The Poor Rogue Trader, Forbes, June 19, 2013.

To study these, and similar episodes, we construct a general misconduct stopping problem in continuous time. A decision maker undertakes an activity which s/he thinks is very likely benign or even beneficial, but s/he acknowledges that there is some small chance that the action is harmful. For example introducing a new drug which the decision maker believes is very likely to be beneficial, or offering some flexibility in the opex/capex distinction to mitigate the need to announce a current cash-flow shortfall anticipated to be short-lived. If the activity is indeed benign then the game continues until the decision maker at some random time exits e.g. through death or merger. But if the truth is that the activity is harmful then at some random time a private signal to that effect arrives, and then after a further random duration a public signal of harm will arrive. If the private signal is received the decision maker will know for sure that the activity is harmful. When the public signal arrives, all will know that the activity is harmful.

The core of the analysis is to understand how an ethical decision maker behaves once s/he receives the private bad-news signal. There are three key forces in our analysis. First, if the decision maker chooses to stop voluntarily the apparently profitable activity then it is a tacit admission that the good or activity is harmful. The decision maker must compensate all prior clients and so a fine is payable whose size is proportional to the number of prior clients. This acts to deter revealing the arrival of bad news. The second force in our model is however the effect of ethics. If the decision maker chooses to ignore the private bad news signal then s/he knows s/he is harming the clients. If the decision maker is ethical then this will reduce his/her utility. The more the decision maker has other-regarding preferences, which is the manner in which we include ethics, then the more the decision maker will dislike continuing with the activity and knowingly harming the clients. The third force in our model is that after the private signal, a public bad-news signal will arrive, though at a random time. Once the public signal arrives the decision maker will be forced to stop and then will be fined for all the people affected. Hence continuing with the known harmful activity creates an increasing liability. The net present value of this activity is reduced by discounting and by the chance that the decision maker will exit through death

or merger. This anticipated future punishment will encourage the decision maker to admit the harmfulness of her activity.

The first and main result of this study is to show that no matter how ethical a decision maker is, formally however much weight is given to the clients' wellbeing, then there exists an *evil time horizon* such that if the private bad news signal arrives after this time has elapsed since the commencement of the business, the decision maker will knowingly prefer to continue with the activity, harming the clients and hiding the news that the activity is harmful. Continuing with a known harmful activity is incredibly painful for the decision maker – the harm knowingly inflicted hurts the decision maker, and s/he knows that eventually a public signal will arrive at which point the punishment will be severe. One mitigating factor is that carrying on might result in escape via death or a merger – but this is not the logic for the result. Rather the key countervailing effect is the realisation that stopping the activity now is a tacit admission of guilt meaning that all the prior clients will have a claim for redress. The longer the activity has continued before the decision maker discovers it is harmful, the greater the punishment that is crystalised for the decision maker if s/he confesses. Eventually, no matter how ethical the decision maker is, confession is worse than concealment and the decision maker will knowingly harm.

This logic helps to explain why any person, no matter how ethical, can find themselves knowingly choosing to do harm. And it also explains why decision makers might bitterly regret their actions, not viewing them as actually right.

The next set of results concerns the sensitivity of the evil time horizon to the parameters of the model. If the decision maker is more ethical by caring more about the clients, then the evil time horizon is pushed out: it takes longer before the decision maker is so committed that s/he will not admit if bad news arrives. A similar effect occurs if the efficacy of public detection improves, so that the public bad-news signal arrives with greater probability after the private bad news signal – the evil time horizon is again pushed out.

However the link between the length of the evil time horizon and the fine rate is more surprising. If the decision maker places sufficient weight on the clients'

well-being (i.e. is ethical) then increasing fines *shortens* the evil time horizon. That is, a high fine makes it more likely the ethical decision maker will be locked into evil and conceal the harmfulness of the activity. The intuition for the result that fines can make bad behaviour worse comes from reflecting on the balance of the three forces on the evil time horizon. If the private bad news signal arrives right on the evil time horizon then the ethical decision maker dislikes equally stopping the activity immediately and incurring the fine versus the combination of hiding the bad news and so putting off a larger fine plus the ongoing pain which will be caused to clients from continuing. If the decision maker is very ethical then most of the reason to reveal and stop the activity comes from the ethicality of the decision maker causing him/her to internalise the harm caused to others, whereas all of the pressure to continue comes from avoiding the fine from confessing. If the fine rises this increases the costs of stopping more than it increases the costs of concealment, and so the decision maker strictly prefers to conceal the harm. That is the evil time horizon moves earlier.

With a new activity whether a given drug or business is good or has unseen side-effects is unknown. Rules are therefore written to manage the worst-case scenario – what should happen if the activity is harmful. We study this setting. Our first result is to establish that welfare is higher if the decision maker is more likely to be ethical. We have established that the decision maker may not admit when s/he finds out an activity is harmful. However more ethical agents have a longer evil time horizon. This means that the probability that bad news will take sufficiently long to arrive that the decision maker will conceal the harm is low. Hence ethical agents are desirable.

We next demonstrate that the optimal fine rate depends upon the ethics of the decision maker. If the decision maker is sufficiently ethical then it is best to have low fines. Low fines make admitting to the arrival of bad news less painful. And the temptation to conceal is weak as an ethical decision maker internalises harm that s/he knowingly causes others. But if the decision maker is unethical then high fines are best for welfare. Concealment will lead to fines in the end as the public bad news signal will arrive eventually and high fines make this eventu-

ality painful for any agent and so push back the evil time horizon for unethical decision makers. Therefore optimal fines depend on the ethics of the decision makers. If the business founders care for their fellow-man then high fines are counterproductive. If the ethics of decision makers is not known for certain, but only probabilistically then we formulate an if and only if test which guarantees that welfare is increased by increasing fines.

Unlike fine levels, there is no ambiguity in a social planner's optimal level of public detection. The faster the public signal of harm follows the private signal the higher welfare will be. Whatever the ethics of decision makers better public detection of harm pushes out the evil time horizon and increases welfare.

Though ethical agents are desirable in that they increase welfare, sadly in an innovation race we show ethical agents are disadvantaged and more likely to lose to unethical competitors. The reason for this is not to do with changes in the length of the evil time horizon. Instead an ethical agent is more concerned before commencing a project with the harm that s/he would do if the new innovation is harmful. This lowers the utility anticipated by an ethical decision maker and so makes unethical decision makers willing to spend more on winning an innovation race. Following similar reasoning we show that projects run by more ethical decision makers are safer on average. The empirical prediction we arrive at is therefore that the average harm perpetrated by new innovations from ethical decision makers will be lower than the average harm perpetrated by unethical decision makers.

If the business opportunity ends, eg a patent expiring, then the evil time horizon continues to exist. No matter how ethical the decision maker, there is a time which is strictly before the termination of the patent such that if the private bad news arrives after this point then the decision mkaer will prefer to conceal the knowledge and knowingly continue to harm through to the expiration of the business opportunity. As above admitting harm crystallises a fine, whereas knowingly harming others to the end and then stopping allows any fine to be at some future point and so subject to discounting. Eventually it is always the case that harming others seems best.

Finally we discuss the policy conclusions of our work. Ethics in this model are the extent to which the harm inflicted on clients is comprehended and internalised by decision makers. Mechanisms in which this can be fostered could be valuable. There are a number of such mechanisms and we discuss them. Finally fines are not substitutes to the efficacy of public detection and so we show that the solution to misconduct is not just to punish more harshly.

2 Literature

It is not hard conceptually to understand why bad people, e.g. those concerned only for their own well-being (*homo-economicus*), might do bad things to others who they care nothing for. This can be explained by a cost-benefit analysis in which the pecuniary benefits of cheating are set against the probability of being caught and punished. The literature captures this trade-off elegantly in a number of contexts: (Tirole, 1996, Sobel, 1985, Becker, 1968). What is much harder to understand is why good people – those who do care about others – do bad things. This care for others can be most simply captured as other-regarding preferences (Fehr and Schmidt, 1999)). If decision makers just don't care *enough* about others (Kajackaite and Gneezy, 2017) then perhaps that is sufficient explanation. But accepting this as the whole story would not explain why some people profess to regret the harm they have willingly wrought – as Mr Yates Jr., formerly of WorldCom, articulated above.

We offer an explanation as to why good people (even very good people) can knowingly choose to do bad things.

There is a significant literature on a related but different question: how do good people come to terms with the fact that they are doing bad things. A key line of research pursued is to explore how decision makers can change their beliefs and, for example, come to view the action as actually a good thing (Bicchieri et al. (2023)). This is explained by arguing that the agent first moves through cognitive dissonance before altering her beliefs so as to come to the view that her actions are

appropriate (Akerlof and Dickens, 1982, Gennaioli et al., 2022, Vitell et al., 2011, Bem, 1967, Lowell, 2012). It might even be that the decision maker refuses to think about information which they dislike, and so in effect behaves as if they are not causing any pain to others (Golman et al., 2017). Unlike this line of research we do not explore motivated beliefs – so our decision makers do not choose to believe that bad is good.

A second line of research is to argue that agents are unable to resist short-term temptation and so find themselves, for example, eating chocolate even though they are on a diet. This literature has usually modelled decision makers as having two selves: a *want self* and a *should self*. The want self over-emphasises short term gains and cannot over-rule the should-self (Bazerman et al. (1998)).[10] This helps to explain why restricting choice sets to an agent can be valuable (Thaler (1980)). If agents have imperfect memory then they cannot recall why they took an action which seems ex post to suggest they are bad (Shu et al. (2011)), and so they can rescue their sense of self-worth by undertaking some good works to invest in their moral capital (Bénabou and Tirole (2011)).

We develop a stopping problem in which a decision maker discovers that his/her activity is harmful and must decide whether to stop the activity or not. Stopping problems can be analysed using dynamic programming (Dixit and Pindyck, 1994). In the model presented signals will arrive at times given by exponential distributions which will afford closed-form solutions amenable to analysis (Benmelech et al., 2010).

Finally, it has been a long-standing target amongst Finance and Economics scholars to include moral reasoning into economic modeling (Arrow (1973), Hausman and McPherson (1993)). Though a majority of the literature has avoided attempts to deal with ethics, there have been others who have made contributions. Some progress has been achieved on why a market might flip to misconduct even when the market participants are ethical: Easley and O'Hara (2023), Thanassoulis (2023). This paper differs by establishing why even the most ethical agents can

[10]Such preferences have been axiomatised and can be captured as preferences over the choice set: Gul and Pesendorfer (2001).

find themselves choosing to commit misconduct. Business Ethics research has a long history in characterising the ingredients which they consider good models of misconduct should contain. The most celebrated of these is perhaps the *Fraud Triangle* (Cressey (1953), Schuchter and Levi (2016)). Through a pioneering series of interviews with white-collar criminals convicted of embezzlement, Cressey (1953) identified three ingredients necessary for wrong-doing to occur: (i) the decision maker must face a problem which s/he deems is non-shareable and desperate. In modern works this is captured as having the incentive. (ii) The individual is in a position to fix the problem – they have the opportunity. And (iii) the ability to rationalise the action and so justify to him/herself why s/he is committing the fraud. The decision maker we study in this paper has all of these ingredients.[11]

3 The Model

A decision maker has decided to undertake a possibly, but s/he believes unlikely to be, harmful activity. Examples are selling a drug after proper – but not exhaustive – testing, booking in advance earnings which are very likely to materialise, or circumventing risk controls to take on a trade which the decision maker is convinced will in any case be realised in his/her favour.

At the point of commencing the decision maker realises the activity might be harmful but places a probability $q \approx 0$ that this is the case. With probability $1 - q$ the decision maker believes the activity is not harmful. Equivalently for our analysis, the decision maker believes that in the not-harmful state any harm is *de minimis*. For example, the Sackler brothers (Oxycontin) very likely did believe they had discovered an important treatment for pain – a view supported by a number of doctors at the time (Keefe, 2021), whereas Mr Burford Yates Jr (WorldCom) was no doubt aware that coding opex as capex the first time was wrong, but he

[11] Relatedly Dupont and Karpoff (2020) argues that the existence of any economic interaction relies on the *Trust Triangle*; trust is created by the existence of (i) ethics on the part of the seller, (ii) laws and sanctions in the event of breaches, and (iii) societal punishment through business reduction or termination in the event that misconduct becomes public. Again, all these critical ingredients are contained in our modelling and analysis.

anticipated that it would be a short-lived intervention and the effects would be immaterial as the good business results he was anticipating would imminently roll in.

Once the decision maker decides to start the activity then the following game unfolds in continuous time. The activity generates a profit flow of π. If the activity is benign (or beneficial) to third parties such as customers, clients, and investors, then we normalise the gain these parties receive to 0. However, if the activity is harmful then it generates a flow of harm to third parties of magnitude $v > \pi$ per unit of time. The decision maker is ethical with parameter $\alpha \geq 0$, which measures the extent of other-regarding preferences in the decision maker's utility function (Fehr and Schmidt, 1999). Thus α measures the proportion of third party harm that the decision maker internalises. If the decision maker knew that the practice was harmful then she would have a flow utility of

$$\pi - \alpha v$$

for each moment of time.

The decision maker discounts the future at rate r, and each period of time dt the agent is 'saved' by merger, death or other forced removal with probability $s dt$. If saved the game ends for the decision maker who receives a termination payoff of 0.

If the truth is that the activity is benign (i.e. not harmful) then third parties suffer no harm and the profit flow π continues until the game ends through merger or decision maker 'death'.

If the truth is that the activity is harmful – e.g. anticipated future revenues will not materialise, or the drug has a very bad side-effect – then a private signal will be received by the decision maker with probability $b\,dt$ over each small time interval dt. The arrival of this private bad-news signal is therefore perfectly revealing to the decision maker.

Suppose that a private bad-news signal is received by the decision maker at time t_0. We assume that a public signal will now inevitably follow, though its

timing is uncertain. After the private bed-news signal a public bad-news signal arrives each period of time dt with probability cdt. The truth that an activity is harmful becomes public via a number of channels. It can be the build-up of individuals who are harmed and who complain to the authorities, as in the case of the many who became addicted to OxyContin (Keefe, 2021); it can be due to academic research uncovering the truth, as in the case of Volkswagen emission cheat devices (Ewing, 2017), or as in the research which connected smoking to cancer;[12] or it could be via whistleblowers as in the case of Boeing's approach to the 737 MAX aeroplane.[13]

If a public signal arrives at calendar time t then the decision maker is 'caught'. S/he must immediately stop the activity and must pay a fine of ϕt. The fine equals a fine rate ϕ times the number of people affected, captured by the total length of time the good has been sold for, t. We discuss the interpretation of the fine shortly. If the decision maker stops the activity voluntarily (i.e. without a public signal) at calendar time t then she is also fined ϕt as stopping an apparently profitable activity is a tacit admission of guilt.

If the agent therefore stops the activity as soon as the private signal is received then her payoff at the point of confession will be $-\phi t_0$. Any delay to do so will grow the fine, and will also increase the suffering of third parties, though the profit flow will continue. The harmfulness of the activity will eventually be revealed, though exactly when is uncertain. The decision maker may escape punishment through merger or death first.

A depiction of the game is presented as Figure 1.

The punishment rate ϕ has at least two interpretations. First ϕ can be seen as modelling the legal liabilities and criminal regime in a country. If a firm or individual causes harm to third parties then each third party has the right to sue and receive appropriate redress. This might be achieved by individual court cases, by the amalgamation of the cases into a class action suit, or by the authorities

[12]The key initial research determining that tobacco is carcinogenic was arguably Doll and Hill (1950).

[13]See *Second Boeing whistleblower dies after raising concerns about 737 MAX*, The Independent, May 4, 2024.

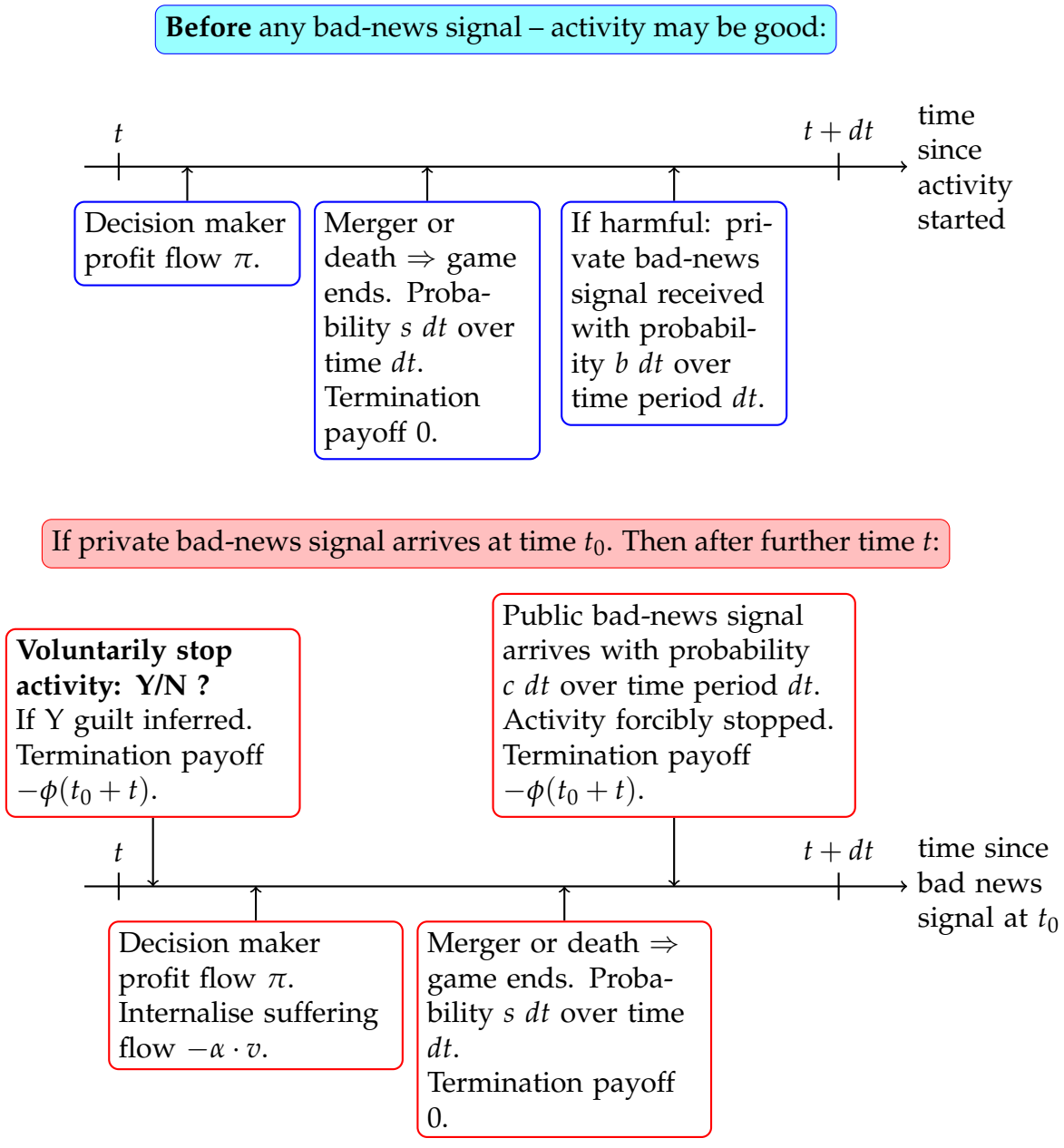


Figure 1: Model timeline

acting on the injured parties' behalf.[14] The decision maker is not protected by limited liability as jail is possible (e.g. Elizabeth Holmes[15]) and the wider family of the perpetrator can also be targeted as a result of the benefit they received from the infraction (e.g. the Sackler family denied protection from bankruptcy for the OxyContin scandal by the US Supreme Court.[16]).

A second interpretation of the punishment rate ϕ is that it captures the shame of being identified publicly as the cause of the clients' suffering. Humans care how they are perceived (Dufwenberg and Dufwenberg, 2018, Abeler et al., 2019) and for many being seen to be money motivated (e.g. choosing to profit from a product which it transpires is harmful) is shameful and lowers utility (Bénabou and Tirole, 2006). Equivalently the shame might arise from the loss of a reputation for being skilled (e.g. Bernie Madoff who had been feted for his investment skill as chair of the Nasdaq stock exchange before the truth of his Ponzi scheme broke.[17]), or the decision maker might gain a reputation as a cheat and a liar which is also undesirable (Gneezy et al., 2018). The fine ϕ captures this interpretation also.

In our benchmark model there is no defined calendar end to the profitability of the product, and so the game has no defined end time. This allows the incentive to wait indefinitely for something to turn up to be studied most starkly. We extend the study to the setting in which the business opportunity has a defined end date, as in the case of a patent, in Section 8.1.

If escape arrives in the form of merger or death then we have modelled the decision maker as exiting the game with a terminal payoff of zero. This is a convenient, but inconsequential assumption. In the case of death – eg the original minds behind the Oxycontin drug, Arthur, Raymond and Mortimer Sackler – this is appropriate. Nonetheless, the merger termination value assumption can be adjusted so that the decision maker never escapes liability for sales that were under his/her control. The results we will present are robust to any such embellishment

[14] See for example Daughety and Reinganum (2005).

[15] Theranos CEO, sentenced to 11 years in jail see link here.

[16] *"Supreme Court Jeopardizes Opioid Deal, Rejecting Protections for Sacklers"* New York Times link.

[17] See *Ex-Nasdaq chair arrested for securities fraud*, CNN money, December 12, 2008.

of the model.[18]

In the core model the fine rate when stopping voluntarily (ϕ) is the same as when stopping is enforced due to the arrival of a public signal that the activity is harmful. In Section 8.2 we extend the model and allow the fine rate to be increased or decreased by ϕ^{extra} if the practice only ceases when the public signal arrives. Authorities often reduce fines to encourage cooperation,[19] but decision makers might feel particular shame if the harm they have been creating is publicly disclosed (e.g. as voiced by Bernie Madoff in the Introduction).

We first solve the core model. All proofs are contained in the Appendix (Appendix A) except for those of the extensions offered in Section 8 which are contained in the Online Appendix (Appendix B).

4 The business evil problem

Theorem 1 *For any level of decision maker ethics, $\alpha \in \mathbb{R}_+$, there exists an* evil time horizon $t_{0*}(\alpha) \geq 0$ *such that if the private bad news signal arrives after $t_{0*}(\alpha)$ then the decision maker will continue with the known harmful activity until s/he is either saved by merger or death or the authorities receive the public bad-news signal causing cessation to be enforced.*

Theorem 1 characterises the decision maker's optimal course of action if she receives the private bad-news signal at time t_0 which confirms that the activity she has started at calendar time 0 is actually harmful. At this point (calendar time t_0), the decision maker has three broad strategies which she could pursue.

The first is confess immediately that it has come to her attention that the activity is harmful and so she will voluntarily stop it. As the decision maker has already been conducting the (now known harmful) activity for time t_0 then she

[18]The termination payoff in the case of merger would then be less than zero as a fine can be expected at some point in the future when the public signal finally arrives. But the present discounted value of this will be more appealing than immediately admitting and paying the fine.

[19]See Financial Times August 13 2024 who write: "*In US capital markets, the wheels of justice turn slowly and grind out exceedingly small fines.*"

will have to pay her fine $\phi \cdot t_0$ and the game ends.

A second course of action is to ignore the bad news and continue with the, now known harmful activity, regardless. This course of action creates a horse-race. It is now a certainty that at some point there will be a public signal confirming the harmfulness of the activity. When this signal arrives then the activity will be forcibly stopped and the decision maker will have to pay a fine proportional to the entire operational life of the activity. However this day of reckoning is avoided if the decision maker dies, either literally or due to a merger or other event relieving him/her of control. (Though prior to escape via merger/death the decision maker will internalise the suffering inflicted on clients.)

The final course of action is a blend of these two extremes. The decision maker could plan to wait until time τ, say, has elapsed since the private bad news signal was received and confess then, accepting the fine at calendar time $t_0 + \tau$. Unless death/merger arrives before this point allowing the decision maker to escape.

The above discussion is captured in the Bellman equation for the decision maker whereby the value of the decision maker who received the private bad-news signal at calendar time t_0 and the decision maker has not admitted that the action is harmful for a further t units of time is:

$$F(t;t_0) = \max \left[\begin{array}{ll} -\phi(t+t_0), & (\pi - \alpha v)dt - \phi(t+t_0)cdt + sdt \cdot 0 \\ & +\frac{1}{1+rdt}(1-sdt)(1-cdt)F(t+dt;t_0) \end{array} \right] \quad (1)$$

The first term in the maximisation is the payoff if the decision maker admits that the action is harmful immediately, i.e. at calendar time $t_0 + t$. The second term is the value which can be expected if the decision maker decides to keep the bad activity going for a further dt units of time at least.

We first describe the proof steps and then explain the intuition for the result.

The first step in the proof is to solve explicitly for what the value function would be if the strategy of the decision maker was not to confess until time $\tau \in [0, \infty)$ elapses after the arrival of the bad news signal. Solving for the value func-

tion in this case is complicated by the fact that the termination value is a function of calendar time, and so the problem is not Markovian (not time independent). This is a crucial feature of the problem as the longer a decision maker has persisted with a harmful activity, the greater the fine that society demands.

The value function of the decision maker who intends to deny there is any problem for τ time after receiving the private bad-news signal before admitting (captured as equation (16)) balances the Poisson processes which govern the arrival of the public news signal and the possibility of merger or death, with the discount rate, the profit flow, and the ethical displeasure of knowingly harming third parties.

The second key step of the proof is to identify that the optimal waiting time before confessing after private bad news (denoted τ above) is either 0, that is not to wait at all and to confess immediately if a private bad news signal arrives, or ∞ i.e. to wait indefinitely and never confess, relying on death or a public order to bring the activity to an end. In other words, waiting for a finite period of time $\tau > 0$ before planning to admit is never optimal for the decision maker. This is established by identifying a sufficient statistic for this model, which is labelled G in the proofs and is given in (17). This sufficient statistic G combines all the variables of the model (ethics, profit, probability of death and discovery etc.). The proof demonstrates that if G is positive then the supremum of the value function is achieved as $\tau \to \infty$, and so never confessing is optimal. If the sufficient statistic G is negative then it is shown that the supremum in the value function is achieved as $\tau \to 0$ so that immediate confession is optimal.

The final step is then to unpack the sufficient statistic G and show that its sign is monotonic in the time at which the private bad-news signal arrives, t_0. This therefore generates a cut-off time t_{0*}, which this work refers to as the *evil time horizon*. If the private bad news-signal arrives before this critical cut off then the decision maker optimally confesses. If the private bad-news signal comes after the critical cut-off t_{0*} then the decision maker, even though she is ethical, never confesses and carries on perpetuating the harmful activity.

Theorem 1 establishes that no matter how ethical the decision maker is, that is no matter how intensely s/he internalises the pain caused to third parties, it is possible for a situation to be created such that the decision maker prefers to continue knowingly causing harm to others rather than admit that the activity s/he is undertaking is harmful.

The intuition for this is as follows. Continuing with the activity causes the decision maker pain as s/he feels for the third parties, and furthermore realises that when the public signal inevitably arrives then the fine will be significant. This is moderated by the fact that it is always possible that the day of reckoning will be avoided if death or merger should come first. Nonetheless, for very ethical people none of this is attractive. However it might be even more unattractive to admit that a practice is harmful if the decision maker has already been pursuing it for a significant period of time. A hefty penalty would have to be paid immediately for damage already done if s/he admits the activity is wrong. Theorem 1 shows that no matter how ethical the decision maker is, there is always an evil time horizon such that bad news arriving after this point will be suppressed.

Theorem 1 provides some insight into how good people can find themselves locked into behaviours they know are bad and greatly regret. The Sackler brothers arguably did believe that OxyContin was a major and safe advance against the widespread problem of pain in the United States (Keefe, 2021). They certainly convinced the FDA of that position. And yet as the evidence of addiction mounted they chose not to stop selling the drug. Similarly Jérôme Kerviel believed that his circumventing risk limits was harmless (profitable for his employer in fact), and yet as evidence came in that his bets would not come good he preferred to double-down rather than confess.

But perhaps the effect captured by Theorem 1 is expressed most poignantly in Bernie Madoff's court transcript. He stated at trial[20]

> *"When I began the Ponzi scheme, I believed it would end shortly and I would be able to extricate myself and my clients from the scheme. However,*

[20]Click here for the Madoff trial transcript.

this proved difficult, and ultimately impossible, and as the years went by I realized that my arrest and this day would inevitably come."

5 The Evil Time Horizon

The proof of Theorem 1 establishes that the evil time horizon is given by:

$$\textit{Evil time horizon } = t_{0*}(\alpha) := \max\left(0, \frac{1}{\phi(r+s)}\left[\frac{c\phi}{r+s+c} + \alpha v - \pi\right]\right) \tag{2}$$

We now interrogate the characteristics of the evil time horizon.

Proposition 2 *The evil time horizon beyond which the ethical agent will knowingly continue to harm* $(t_{0*}(\alpha))$ *satisfies:*

1. *More ethical decision makers have a larger evil time horizon, taking longer before being trapped into evil:*
$$\frac{\partial}{\partial \alpha} t_{0*}(\alpha) \geq 0$$

2. *If punishments are increased then the evil time horizon responds as follows:*
$$\frac{\partial}{\partial \phi} t_{0*}(\alpha) \begin{cases} < 0 & \text{if } \alpha v > \pi \text{ (most ethical)} \\ \geq 0 & \text{if } \alpha v < \pi \text{ (least ethical)} \end{cases}$$
so greater punishment implies the most ethical agents are trapped into evil sooner (have a shorter evil time horizon).

3. *Better public detection* $(c \uparrow)$ *causes the decision maker to have a larger evil time horizon (take longer before being trapped into evil):*
$$\frac{\partial}{\partial c} t_{0*}(\alpha) \geq 0.$$

The comparative statics described in Proposition 2 are depicted in Figure 2. We now explain the intuition.

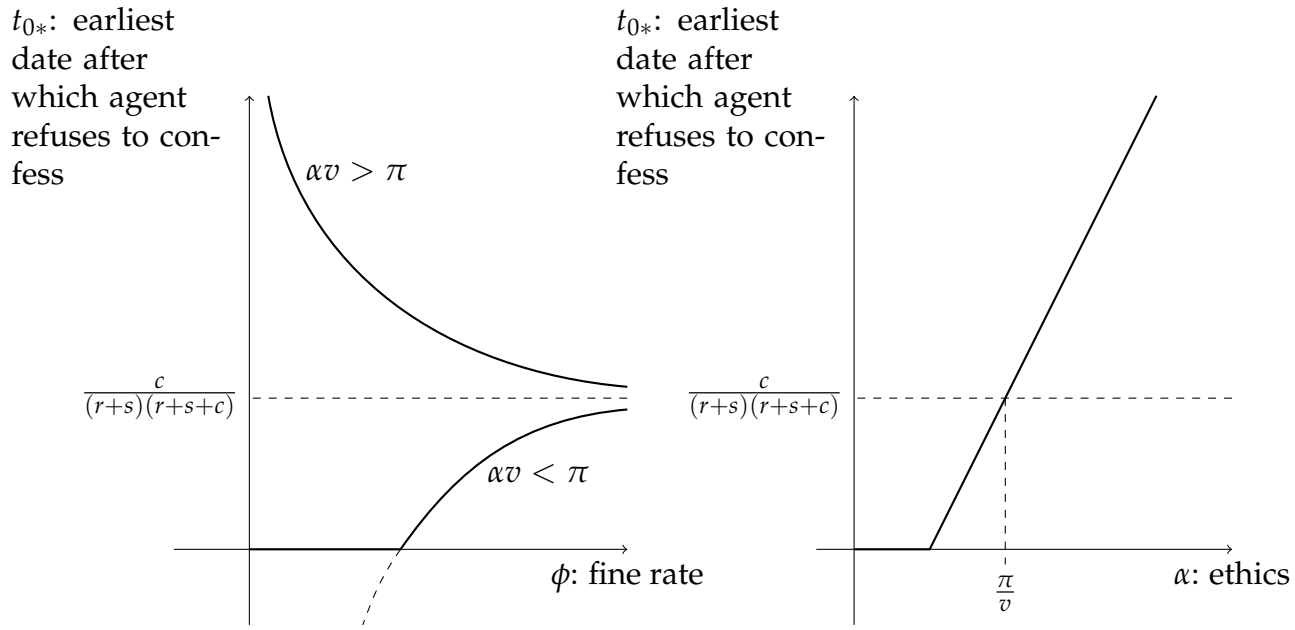


Figure 2: The last point at which agent is willing to confess: t_{0*}.
Notes: The evil time horizon is the critical time after which the agent will not confess that the activity is harmful. The evil time horizon is labelled $t_{0*}(\alpha)$ and is given in Equation (2). The left hand graph plots t_{0*} with respect to the fine rate ϕ. The right hand graph plots t_{0*} with respect to ethics α. The graphs depict the first two parts of Proposition 2.

The evil time horizon is the point in time at which a decision maker is indifferent between confessing to the activity being harmful and paying a fine proportional to the length of time the activity has already been going on, versus denying the harm and continuing with the bad activity, knowingly harming third parties, courting an even larger fine on the day of reckoning when the public signal arrives, but with the hope of deliverance via death or merger. If the decision maker is made more ethical ($\alpha \uparrow$) then the harm inflicted on third parties from a strategy of denial weighs more heavily in the decision maker's utility function. So denial is less appealing and the decision maker strictly prefers to confess. It follows that the evil time horizon moves out.

The intuition for the third result is similar. If public detection improves ($c \uparrow$) then the expected date of the day of reckoning when the public signal will arrive moves forward in time, and so the probability that the decision maker will be delivered by death or merger before then shrinks. This makes the strategy of denial less appealing so that the decision maker would strictly prefer to confess. It again follows that the evil time horizon moves out.

The intuition for the fact that increasing penalties brings forward the evil time horizon and so brings forward the date of lock-into evil for the most ethical is more subtle. Recall that the evil time horizon t_{0*} is the moment in time such that if the private signal arrives at this moment then the decision maker will be indifferent between confessing that the practice is confirmed bad and paying her fine, versus the alternative of continuing with the practice in the hope of salvation from death or a merger before the harm becomes public. The option of deferring indefinitely has payoff composed of two parts: (i) the npv of the fine which may well become payable if the public signal is received before the agent can exit through merger (or death); plus (ii) the flow payoff of the difference between the profit earned and the harm the agent knows she is causing ($\pi - \alpha v$). This can be seen mathematically as from the proof of Theorem 1:

$$G|_{t_{0*}} = 0 \;\Leftrightarrow\; \underbrace{\phi(r+s)t_{0*}}_{\substack{\text{fine from}\\ \text{confessing}\\ \text{immediately}}} = \underbrace{\phi\frac{c}{r+s+c} + \alpha v - \pi}_{\substack{\text{payoff rate}\\ \text{from continuing}}} \tag{3}$$

The time t_{0*} occurs when these twin evils are in balance. If $\alpha v > \pi$ then the fine incurred from immediate confession must be greater than the fine npv component from indefinite waiting as the flow of utility from denial is itself negative (as the ethics cost exceeds the profit gain for every moment of denial). This implies that the sensitivity to the fine of immediately confessing must be larger than the sen-

sitivity to the fine of the payoff from denial.[21] It follows that increasing the fine rate ϕ causes the payoff from confessing to become more negative than the payoff from continuation. The decision maker would therefore prefer denial and hope for future deliverance. It follows that the evil time horizon, the lock-in moment, t_{0*}, moves forward in time.

A further, I think surprising, result is the robustness of the evil time horizon even if the decision maker is an employee and does not benefit from the profits made by the activity. This was the case with the instigator of the WorldCom scandal, Mr Burford Yates Jr., as described above.

Corollary 3 *Suppose that the decision maker does not internalise any of the profits made from sale then there remains a positive evil time horizon given by*

$$t_{0*}^{employee} = \frac{\alpha v}{\phi(r+s)} + \frac{c}{(r+s)(r+s+c)} \in (0, \infty).$$

Proof. Set $\pi = 0$ in the expression for t_{0*} given in (2). ■

Key to Corollary 3 is that the employee experiences pain from being found to have harmed his/her clients, captured by the fine term $\phi > 0$. This fine need not be monetary. As described in the introduction the fine could capture the negative personal utility which would follow if the authorities seek a custodial sentence, as in the case of Jérôme Kerviel and Societe Generale. It could also include the shame which the decision maker feels once their transgression becomes public, as Mr Yates Jr, the author of the WorldCom accounting scandal, voiced.[22]

The intuition to Corollary 3 is then as follows. If a significant amount of time elapses before the private bad news signal arrives, then the decision maker must compare the payoff from confessing against that from continuing and knowingly harming clients. Both confession and continuing hold the promise of fines, from

[21] Mathematically, using (3) above

$$\alpha v - \pi > 0 \;\Rightarrow\; \frac{c}{r+s+c} < (r+s)t_{0*}.$$

[22] See footnote 5.

punishment or shame. But in the case of continuing, even though there is no offsetting profit, and there is the further cost of knowingly harming, the fine element is deferred for an uncertain period of time. The effect of discounting on this fine will ensure that continuing is better than confessing if enough time has elapsed since the activity started. That is, we again have an evil time horizon.

We conclude this section by observing that even if the decision maker is completely unethical, $\alpha = 0$ so that the we revert back to *homo economicus*, then there can still be a positive evil time horizon. That is, even decision makers who care nothing for others can find themselves preferring not to continue with a harmful, though profitable, activity. Though this only happens soon in the life of the business:

Corollary 4 *Consider a completely unethical decision maker* $(\alpha \equiv 0)$*. Such a decision maker will have a strictly positive evil time horizon if:*

$$t_{0*}(0) > 0 \;\Leftrightarrow\; \frac{c\phi}{r+s+c} > \pi.$$

Proof. Set $\alpha = 0$ in the expression for t_{0*} given in (2). ■

If the private bad news signal arrives early enough then the fine from confessing will be negligible. The benefit of continuing is the profit, but the cost is the fine which will be being accrued, payable when the public signal finally arrives. If this fine is large, or if the public signal arrives with a greater frequency, then the fine term will dominate. In this case the decision maker would rather confess and stop the harmful activity. However, after the evil time horizon one must add to this decision calculus the fine payable upon confession. As this grows larger it encourages the unethical agent to continue with the (known harmful) activity.

6 Welfare if Activity is Harmful

Whether the activity is harmful is not known for certain when the activity is started. The decision maker places a probability of q on the activity being harm-

ful. However views likely differ between the decision maker and other interested parties.[23] Regulators therefore often design policies and sanctions to manage the worst case scenario. For example, financial regulation in the UK was redesigned after the Global Financial Crisis to reduce the impact of future crises on the UK economy, not to manage any tradeoffs between bank risk and financial sector competitiveness.[24] We therefore consider welfare for the case in which the activity is harmful. We assume that the social planner and the decision maker share the same discount rate r.

Proposition 5 *If the activity is actually harmful then the expected welfare when the activity begins is:*

$$W(\alpha, \phi, c) = -\frac{v - \pi}{r + s + b}\left[1 + \frac{b}{r + s + c}e^{-(r+s+b)t_{0*}}\right],$$

where t_{0*} is given in (2)*.*

Proposition 5 provides a simple and tractable expression for the expected level of welfare if a harmful activity is started. Proposition 5 allows for the uncertainty as to how long the harmful activity will survive. This survival time depends on exogenous events such as the likelihood that the decision maker will die or his/her firm will be bought out (denoted by s). The lifespan of the activity also depends upon how rapidly the decision maker will find out that the activity is harmful (b), whether this epiphany occurs before or after the evil time horizon (t_{0*}), and how rapidly the public confirmation that the activity is harmful will arrive (c).

The method of proof of Proposition 5 is as follows. Once a harmful activity starts, whether it will be terminated voluntarily or not depends upon whether the private bad news signal arrives before or after the evil time horizon. The bad news signal is a Poisson process. It follows that the arrival time of the signal is distributed according to an Exponential distribution. We can therefore consider in turn the welfare generated if the arrival of the private bad-news signal is before

[23] E.g. decision makers typically suffer from overoptimism bias (Malmendier and Tate, 2005).
[24] See Vickers (2011) and the aims of the suite of regulatory reforms listed in Box 1.1.

t_{0*}, versus if it is after, and then integrate over all these possibilities using the appropriate probability weight from the Exponential.

If the private bad news signal will arrive after the evil time horizon, the decision maker will hide that news and not confess. From the moment that the bad news signal arrives the social planner can anticipate how welfare will evolve: the (negative) flow $\pi - v$ will continue until either the public bad news signal arrives or the decision maker dies (merger/death). The welfare value function can therefore be characterised and its value solved for. This represents the terminal value of welfare at the point the private bad news signal is received.

The social planner can now discount this terminal welfare back to the present – the beginning of the business activity. This discounting must again take account of the possibility of decision maker exit through death. This discounting requires the welfare function to be characterised by an appropriate ordinary differential equation through time, solved for, and then value matched with the terminal welfare value at the point the private signal arrives.

Repeating this process for all possible arrival times of the private bad-news signal, and integrating using the Exponential distribution of the arrival times, allows us to derive the (much simplified) expression in Proposition 5.

We can now characterise how a social planner could increase welfare, in the case that the activity is harmful.

6.1 Ethics are desirable

In this section we show that welfare if the activity is harmful is increased (is made less negative) if the decision maker is more likely to be ethical.

Suppose ethics are distributed in the population of decision makers according to one of the cumulative distributions $\{H^{ethical}(\alpha), H^{unethical}(\alpha)\}$, such that $H^{ethical}(\alpha)$ first-order stochastically dominates $H^{unethical}(\alpha)$:

$$H^{ethical}(\alpha) \leq H^{unethical}(\alpha) \quad \forall \alpha \in \mathbb{R}_+, \tag{4}$$

with strict inequality for some values of α. That is, if decision makers are drawn from $H^{ethical}$ the probability they have a high sensitivity to others' utility and pain is greater (α is more likely to be large) than if decision makers are drawn from $H^{unethical}$.

Theorem 6 *If the activity is harmful, then expected welfare is higher if the ethics of the decision maker are distributed according to the more ethical $H^{ethical}(\alpha)$ than if ethics are distributed according to the less ethical $H^{unethical}(\alpha)$.*

Theorem 6 is proved by taking the integral of welfare values, given in Proposition 5, over the possible levels of other-regarding preferences, α. It is apparent from Proposition 5 that welfare is higher (less negative) if the evil time horizon t_{0*} is greater. The intuition for this is that the social planner wishes to maximise the probability that the decision maker stops the activity as soon as s/he finds out that it is harmful. This only happens if the bad news signal arrives before the decision maker's evil time horizon. The larger the evil time horizon, the more likely that is. The result then follows from the result that the evil time horizon grows in the level of decision maker ethics (established as the first result of Proposition 2).

Theorem 6 suggests that welfare is improved if decision maker ethics are good. There are a number of ways in which policy makers have sought to improve business peoples' ethics. For example, professional certification frequently, if not always, requires some learning in professional ethics (Patel, 2014); regulators can interview decision makers to ascertain if they believe they are fit and proper (Leiper, 2017); firms can choose to implement (or to withdraw) ethics training;[25] decision makers can choose to spend time with customers (Hult, 1929).[26] Theorem 6 suggests that to the extent that such activities encourage decision makers to see things from the viewpoint of the people their decisions affect, the better.

The benefit of better ethics amongst decision makers is not monotonic, and is concentrated amongst intermediate levels of ethics, as depicted in Figure 3. At

[25] See for example *Big tech companies cut AI ethics staff, raising safety concerns*, Financial Times, March 29, 2023

[26] See also *CEOs should get out of the board room spend time on the shop floor*, 2 October 2018, Cityam, with link here.

very low levels of ethics the evil time horizon can be identically zero (depending on the relation between fines and profits in Corollary 4), in which case small improvements in ethics have no impact on welfare. At the other extreme, if the decision maker is incredibly ethical then their evil time horizon will be very large. The probability that the private bad news signal arrives after this point becomes very small. Thus the likelihood is that the decision maker will have discovered the activity is harmful before their evil time horizon. So further improvements in ethics are unlikely to make a significant difference to expected welfare. At intermediate levels of ethics however, improvements in ethics have the biggest effect.

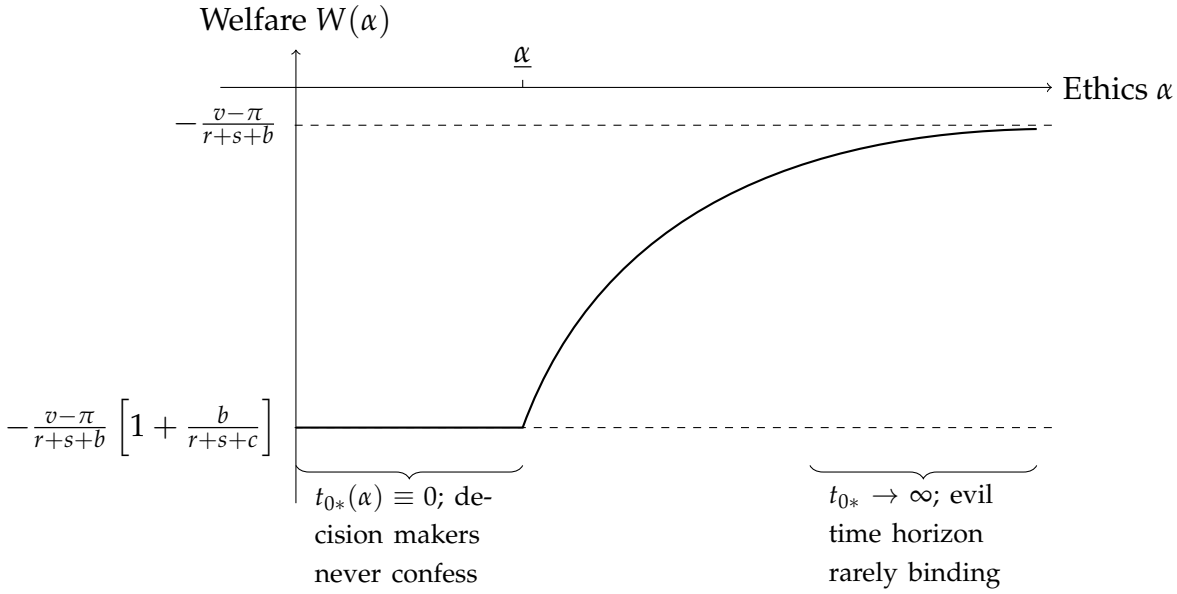


Figure 3: Welfare is not strictly monotonic in ethics

Notes: Welfare is given in Proposition 5 for the case in which the product is actually harmful. The ethics level $\underline{\alpha} := \frac{1}{v}\left[\pi - \frac{c\phi}{r+s+c}\right]$ is the highest level of ethics such that the evil time horizon is 0. It is evident that increasing decision maker ethics is most valuable at intermediate levels of ethics.

6.2 Optimal fines depend on ethics

High fines (ϕ) have opposing effects. On the one hand as the fine will be high when the public signal arrives a decision maker who discovers that the activity is harmful will be encouraged to confess and avoid this large cost. But high fines also make confessing immediately more expensive and so deter this. In Section 8.2 we will consider the case where the decision maker feels (or is fined) differently depending upon whether s/he or the authorities inform the public that the activity is harmful. Here we consider the model of one fine rate irrespective of whether the activity stopped voluntarily or was forced.

Suppose that the fine rate ϕ is restricted to lie in a range $\phi \in [\underline{\phi}, \overline{\phi}]$. The lower bound is motivated by noting that third parties who are harmed have rights to redress – we assume these rights cannot be cancelled by a policy maker. The upper bound is motivated by the fact that the law typically requires fines to be commensurate with the harm inflicted.[27] In particular excessive fines are not permissible legally.

We establish that:

Theorem 7 *Suppose that the fine rate ϕ is restricted to lie in the range $\phi \in [\underline{\phi}, \overline{\phi}]$. Suppose that the activity is harmful.*

1. *If the decision maker is sufficiently ethical, $\alpha > \frac{\pi}{v}$, then the welfare maximising fine is as small as possible: $\phi^* = \underline{\phi}$.*

2. *If the decision maker is sufficiently unethical, $\alpha < \frac{\pi}{v}$, then the welfare maximising fine is as large as possible: $\phi^* = \overline{\phi}$.*

3. *If the ethics of the decision maker are not known with certainty, but distributed according to density function $h(\alpha)$, $\alpha \in \mathbb{R}_+$ then the expected value of welfare is*

[27] If the fine captures shame then we note that the disutility decision makers internalise from harm is often proportional to the amount of harm inflicted (Thanassoulis, 2023).

increased by an increase in fines if and only if

$$\frac{M_h'(-v\psi)}{M_h(-v\psi)} < \frac{\pi}{v} \tag{5}$$

$$\textit{where } \psi := \frac{1}{\phi}\left(1 + \frac{b}{r+s}\right). \tag{6}$$

and $M_h(x) = \int_{\alpha=0}^{\infty} e^{\alpha x} h(\alpha) d\alpha$ *is the moment generating function of* h.

Theorem 7 is proved by establishing the comparative static of the rate of change of expected welfare, given in Proposition 5, with respect to the fine rate ϕ. The expected welfare depends tightly on whether the decision maker will confess when s/he receives the private bad news signal. In turn this depends upon the evil time horizon, t_{0*}. And Proposition 2 established that the sensitivity of the evil time horizon to the fine rate depended on the ethics of the decision maker.

Theorem 7 shows that optimal sanctions depend upon how ethical the policy maker considers the decision maker to be. If the decision maker is ethical then fines create a lock-in problem in which even ethical decision makers can fall into the evil of denial whereby they continue with an activity that they dislike to avoid the even worse fine from confessing. It follows that if the decision maker is ethical (specifically $\alpha > \pi/v$) then the optimal fine is as low as possible commensurate with the rights of those already harmed, $\underline{\phi}$. With ethical decision makers this pushes the evil time horizon as far out as possible. But if the decision maker is unethical only the fine can force him/her to confess, and then only if the bad news comes quickly enough. As before the optimal fine pushes the evil time horizon as far out as possible, but with an unethical decision maker this requires a fine rate which is as high as possible, as confirmed by the second result of Proposition 2.

From Theorem 7 we have established that if there is a known level of ethics in the decision maker (α) then if $\alpha < \frac{\pi}{v}$ it follows that it is optimal for welfare if the fine rate is increased up to the maximum, and not otherwise. If however the policy maker is uncertain of the ethics of the decision maker then a different condition governs when fines should be as high as possible. Condition (5) can be

interpreted as saying that the expected value of α should replace the known value of α in Theorem 7, plus a correction. To see this note that if $v\psi \approx 0$ then the left side of condition (5) is:

$$\frac{M_h'(-v\psi)}{M_h(-v\psi)} \approx \frac{M_H'(0)}{M_H(0)} = \frac{E_H(\alpha)}{1}$$

by the definition of the moment generating function.

Example 8 *Suppose that ethics α are χ^2 distributed in the population with $\bar{\alpha}$ degrees of freedom. The expected level of ethics is therefore $\bar{\alpha}$. The moment generating function in this case is given by*

$$M_h(x) = \frac{1}{(1-2x)^{\bar{\alpha}/2}} \text{ valid for } x < \frac{1}{2}.$$

The ratio of the derivative of the moment generating function to itself is:

$$\frac{M_h'(x)}{M_h(x)} = \frac{\bar{\alpha}}{1-2x}.$$

It follows from Theorem 7 that the fine rate ϕ for the activity when it is discovered to be harmful should be increased if the expected level of ethics satisfies:

$$E(ethics) = \bar{\alpha} < \frac{\pi}{v} + \underbrace{\frac{2\pi}{\phi}\left(1 + \frac{b}{r+s}\right)}_{\ddagger}.$$

The term $(\ddagger)$ corrects for the fact that in the χ^2 distribution a large measure of agents have ethics clustered around 0. Thus it is optimal to increase fines even if the average level of ethics is π/v.

A determination of whether damages to clients hurt by business innovations, and sanctions on white collar crime should be more severe cannot therefore be answered without understanding how ethical agents are. Without a view on the latter question increases to sanctions, for example, can make the outcome when

inventions are harmful worse.

6.3 Better public scrutiny skill increases welfare

The change in welfare if public detection increases is not ambiguous, as it was in the case of fines:

Proposition 9 *Suppose the activity is harmful. Welfare is increasing in the quality of the public signal of harmfulness:*

$$\frac{\partial W}{\partial c} > 0$$

The proof of Proposition 9 follows by differentiating the welfare condition (Proposition 5) with respect to c. Welfare is closely tied to whether the decision maker will be honest and confess when s/he receives the private signal that the activity is harmful. The farther out the evil time horizon is, the better for welfare. Increasing the speed with which the public signal follows the private signal lowers the chance a decision maker will have of escape via merger/death, and it brings forward the expected day of reckoning. Hiding a bad news signal therefore becomes less appealing. It follows that the evil time horizon is pushed out and welfare up.

It is immediate upon comparison of Proposition 9 and Theorem 7 that fines and public detection cannot, in general, substitute for each other. In particular if decision makers are ethical then a reduction in Welfare caused by a reduction in the capacity of the authorities to bring harmful activities to light ($c \searrow$) cannot be remedied by an increase in the penalty decision makers face upon discovery.

7 Entry: ethical agents are disadvantaged, but safer

In this section we study the entry decision of the decision maker. We ask whether in any entry race, ethical or unethical, decision makers are most likely to win. And then we ask how the average quality of the products launched will be related to the ethics of the decision maker.

7.1 Ethical agents are disadvantaged in entry races

Suppose that two decision makers are competing for the right to launch an innovation. This could be a battle for a patent on a drug, or it could be a battle for regulatory approval of a new financial product. What are the characteristics of the decision maker most likely to win this race? This question has been asked in multiple forms in the past. A seminal example is Gilbert and Newbery (1982) who ask if an incumbent firm or an entrant are most likely to win a patent race. The question is answered by identifying the expected value of the activity to each party and interpreting this as a measure of the maximum amount which the decision maker would spend to win. The spending could be monetary (e.g. R&D expenses), or it could be an effort cost (e.g. to persuade a supervisor). We follow the Gilbert and Newbery (1982) approach.

Let us denote the agent's value for the activity if it is harmful as V^{bad}, and V^{good} if the activity is benign. The agent is willing to expend a value $\bar{V}$ to commence the project where:

$$\bar{V}(\alpha) := qV^{bad} + (1-q)V^{good}. \tag{7}$$

Theorem 10 *Suppose decision maker* 1 *is less ethical than decision maker* 2 $(\alpha_1 < \alpha_2)$ *but that the decision makers have the same subjective probability that the activity is harmful,* q*. It follows that decision maker* 1 *(unethical) values the activity higher than decision maker* 2 *(ethical):*

$$\alpha_1 < \alpha_2 \;\Rightarrow\; \bar{V}(\alpha_1) > \bar{V}(\alpha_2).$$

And so the unethical agent is most likely to win an entry race.

The key step in proving Theorem 10 is to establish the value which a decision maker can expect should the product they are launching turn out to be harmful. The method by which this is done parallels the manner in which welfare was established in Proposition 5. The differences are that the decision maker internalises a multiple of third parties' harm (α), and the decision maker is exposed to the sanction regime (fine rate ϕ) which is not part of the welfare calculation. In both cases the proof strategy is to integrate over all the possible arrival times of the

private bad news signal. And for each possible arrival time evaluate the value the decision maker can anticipate by solving the o.d.e. formed by the evoluation of the value function.

Theorem 10 establishes that unethical agents are willing to expend the greatest resources to launch the activity. This is despite the fact that ethical agents have a greater willingness to confess to dangerous practices earlier and so avoid the largest fines.

The intuition behind Theorem 10 is related to the envelope theorem. If we were to increase an agent's level of ethics then she would be less vulnerable to being locked-into evil and be more willing to confess if and when the bad-news private signal is received. As the critical lock-in, the evil time horizon t_{0*}, is selected as the time when the agent is indifferent between confessing and not, the overall change in utility is small. However, an increase in ethics causes ethical agents to be more pained by the harm they do if the action is harmful. This concern is absent from unethical agents. As ethics become more intense, this harm becomes more salient and so lowers the expected value of the entire innovation.

It follows that the very same ethics which causes welfare to rise as agents are more likely to confess, causes ethical agents to be reluctant to launch the activity at all. The model predicts therefore that less ethical agents would win a contest to launch this new activity.

7.2 Ethical agents are safest

We now consider the joint distribution between ethics and harmfulness of products launched. To do this we build on Theorem 10 and so generate an empirical hypothesis concerning the expected quality of products introduced as a function of decision maker ethics.

Label the distribution of ethics in the population of decision makers as $H(\alpha)$. Label the distribution of subjective probabilities of harm which can be held by decision makers concerning possible projects as $J(q)$. Recall q is the probability that the decision maker believes the project is harmful. Let us suppose that deci-

sion makers are randomly assigned to possible projects and they then each choose whether to launch their project. Entry is assumed to cost a fixed amount: κ.

We write $\mathcal{E}_q(\alpha)$ for the expected probability of harm in the population of launched projects run by decision makers with ethics α:

Proposition 11 *Projects run by more ethical decision makers are safer on average:*

$$\frac{d}{d\alpha}\mathcal{E}_q(\alpha) < 0$$

Proposition 11 is proved by identifying the maximum probability of harm, q, which a decision maker will tolerate and still enter, if entry costs a given amount (in the proof denoted κ). In the proof of Theorem 10 we established that decision makers with higher ethics anticipated a reduced value (more negative) of launching a product in the state of the world in which the project is bad. Therefore ethical decision makers see a lower value to entry overall than unethical ones. It follows that the critical anticipated probability of harm above which entry does not happen (denoted $\bar{q}(\alpha)$ in the proof) is lower for more ethical agents. And so a simple integral argument establishes the result that conditional on entry, more ethical agents have products which are less likely to be harmful.

Proposition 11 is, in principle, testable. The statistics on misconduct are collected by authorities and are public as clients seek redress. Measuring the ethics of decision makers can be proxied by measuring their prosociality, more colloquially referred to as their kindness. How to measure this is a subject of ongoing research, though a number of survey approaches have been proposed (Youngs et al. (2023), Wallace and Thurman (2018)).

8 Extensions

8.1 Determined end date – e.g. Patents

In the model we have studied the business activity did not have a defined end date. The evil time horizon result (Theorem 1) therefore abstracted from any end-point effects. In this section we explore business opportunities which only exist over a restricted time period, such as patents. We return to the question of whether it is ever possible for an ethical decision maker to choose to knowingly harm clients. We find that the evil time horizon result is robust.

In this section we assume that the business opportunity is only available until calendar time T. After this time the decision maker receives no further profits and does no further harm as the good is no longer sold. However the rest of the model is unchanged: the decision maker discounts the future, they may at any point be 'saved' by merger or death according to a Poisson process, and if private bad news has arrived then public bad news will follow according to a Poisson process with rate c. If public bad news arrives then the decision maker is fined proportional to the time the business was in operation; hence the fine is $-\phi T$.

Proposition 12 *For all decision makers, a unique evil time horizon $t_{0*}(\alpha) < T$, shorter than the length of the business opportunity, exists.*

Proof. See the Online Appendix. ■

It follows that even the most ethical decision makers will hide private bad news if the end of the business opportunity, or patent, is sufficiently near. To understand why consider an ethical decision maker receiving private bad news a short moment before the patent or business opportunity expires. If s/he admits to the bad news now then a fine will be payable immediately. If the decision maker hides the news and continues with the activity until the patent expires then a public bad news signal will come to light at some point – but at that point the same nominal fine will be payable as no-one further will have been harmed. Due to discounting this is more appealing and so the decision maker will hide the bad

news. It follows that there is always an evil time horizon, for any level of ethics, which lies strictly before the end of the patent.

The proof uses the same steps as those used for Theorem 1. The main difficulty is to establish that upon receipt of the private bad news, only immediate confession or waiting until the business opportunity expires can be optimal. This is done by establishing that the value function of the decision maker is quasi-convex in the length of time s/he plans to wait after the arrival of the private bad news signal. The second step is to note that there is a jump in the value function between waiting until just before the patent expires versus allowing the patent to expire. The value of waiting to the end is increased as any fine only arrives at some point in the future, and so the discounting described in the intuition above increases the value function. A comparison of the payoff between confession versus hiding the bad news signal to the end then yields the result.

8.2 Penalty flexes on voluntary versus involuntary revelation

Suppose that the authorities can adjust the fine payable when they seek to enforce cessation after the public bad-news signal. Label the adjustment ϕ^{extra} so that the total fine paid if the activity only stops when the public signal is revealed after calendar time t is:

$$\text{Fine} = (\phi + \phi^{extra})t$$

We consider three possible fine regimes.

1. Define **leniency** as the case

$$\phi^{extra} < 0 \quad \Rightarrow \quad (\phi + \phi^{extra}) < \phi.$$

   This is the case is which the authorities offer the decision maker an inducement to cease the activity upon the arrival of the public bad news. Leniency regimes are common: the UK and US offer leniency in cartel cases,[28] the

[28] See Gov.uk website link, and Justice.gov website link.

SEC is accused of demanding small fines,[29] and plea deals are a feature of US justice.[30]

2. Define **punishment** as the case

$$0 < \phi^{extra} < \left(\frac{r+s}{c}\right)\phi \tag{8}$$

Punishment regimes increase the fine rate if cessation is forced, but they can be evaded if the decision maker can discover when a public signal is about to arrive, and confess just before. Therefore if the decision maker can pre-empt public enforcement by, for example, monitoring relevant publications, monitoring whistleblowers, or by frustrating the authority's work then the punishment fine is of limited practical relevance.

3. Define a **punitive** regime as one in which the extra fine is disproportionately large:

$$\phi^{extra} > \left(\frac{r+s}{c}\right)\phi \tag{9}$$

This raises the fine payable by the decision maker on the arrival of public bad news significantly. However to the extent that such fines exceed the degree of the offence they are illegal – see for example the US Supreme Court. [31] In addition the punitive regime shares all the drawbacks of enforceability with the punishment regime.

We find that in the punishment and leniency regimes our results are unchanged. But a significant change to equilibrium behaviour occurs if punitive fine regimes can be enforced:

Theorem 13 *Suppose that the authorities levy an extra fine ϕ^{extra} upon public revelation that the activity is harmful, in addition to the stopping fine rate ϕ.*

[29] See footnote 19.

[30] See for example Caroline Ellison gets 2-year prison sentence for FTX fraud.

[31] See for example No. 23-175, Supreme Court of the United States pp3-4, in which it is argued disproportionate sanctions violate the Eight Amendment of the US Constitution.

1. *Theorem 1 applies in the* ***leniency****, and* ***punishment*** *regimes:*
   for any level of decision maker ethics, $\alpha \in \mathbb{R}_+$*, there exists an* evil time horizon $0 \leq t_{0*}(\alpha) < \infty$ *such that if the private bad news signal arrives after* $t_{0*}(\alpha)$ *then the decision maker will continue with the known harmful activity until s/he is either saved by merger or death or the authorities receive the public bad-news signal causing cessation to be enforced.*

2. *In the* ***punitive*** *fine regime*

   (a) *A decision maker with ethics such that*

   $$\alpha v \geq \pi - \phi \tag{10}$$

   is honest and admits the arrival of private bad news immediately.

   (b) *Otherwise there exists a* confession time horizon $0 < \mathcal{T}_{0*}(\alpha) < \infty$ *such that if the private bad news signal arrives before* $\mathcal{T}_{0*}(\alpha)$ *then the decision maker continues the activity regardless (knowing it is harmful) until calendar time* $\mathcal{T}_{0*}(\alpha)$ *before confessing. The decision maker confesses immediately and stops the activity if a private bad news signal arrives after* $\mathcal{T}_{0*}(\alpha)$*.*

Proof. See the Online Appendix. ■

Theorem 13 is proved using the techniques we have established in Theorem 1 and in Proposition 12. In each case the value functions upon receiving the private bad news signal are established if the decision maker were to confess, to hide indefinitely, or to hide for a planned period and then admit. In the case of the punitive fine regime only, the latter can be optimal.

The intuition for the evil time horizon in the leniency and punishment regimes has been explained already. We can explain why there is a period of time before confession in the punitive regime. Suppose that the private bad-news signal arrives the moment that the decision maker starts trading, i.e. at $t = 0$ calendar time. No fines have yet been incurred. Trading for at least time dt and then stopping generates an extra fine $-\phi dt$. The effect of discounting or being caught is

of order dt^2 and so can be set aside. In addition, trading despite knowing the activity is harmful generates a utility flow of $(\pi - \alpha v)dt$. Therefore if the decision maker is unethical enough that (10) does not hold, lying at least initially, is value increasing.

Now suppose that instead of waiting for calendar time $\mathcal{T}$ to confess, the decision maker decides to wait for a further time dt (that is to wait to calendar time $\mathcal{T} + dt$) before confessing. This delays the voluntary fine allowing for discounting, and the agent might be saved with probability sdt. Therefore the decision maker can expect to gain $(r + s)dt \cdot \phi\mathcal{T}$. But if the public bad news signal arrives in that interval, happens with probability cdt, then the decision maker will be faced with an extra fine, so overall the risk is of an extra cost of $cdt \cdot \phi^{extra}\mathcal{T}$. In the punitive regime we see that waiting beyond time $\mathcal{T}$ to confess lowers value overall as from condition (9):

$$(r + s)\phi - c\phi^{extra} < 0.$$

Hence the value function is declining as $\mathcal{T} \to \infty$. Therefore no decision maker finds it optimal to lie forever in the punitive regime. We therefore establish that the optimal lying period in the punitive regime is interior, more than zero and less than infinity, which gives the intuition to Theorem 13.

If the authorities can increase the fine payable by the decision maker on the occasion that the authorities force cessation then the evil time horizon result continues to apply if the fine increase is not too large (punishment or leniency regime). But if the extra fine could be made disproportionately large (the punitive regime) then the model predicts a confession time horizon. In this case no decision maker would ever hide harmful practices for ever. The case of punitive punishment with involuntary revelation is, in my opinion, unlikely to hold in reality – as decision makers may be able to anticipate when public news will arrive, as agencies might prefer to offer leniency in return for cooperation, and as the shame attached to wrongdoing is not strongly related to how the news enters the public domain.[32]

[32]Further the examples of the introduction, such as Bernie Madoff and the Sackler family, suggest that some decision makers really did intend to hide the knowledge that they were harming others indefinitely.

9 Policy Implications and Conclusion

In this paper we have developed a game theoretic analysis to study why ethical decision makers might choose to knowingly harm others. The premise of the analysis is that the decision maker may discover that a practice or good s/he thought was benign is actually harmful. However admitting to this fact means compensating those already adversely affected. Carrying on however will create even more people to compensate when the truth is eventually revealed, and will hurt the decision maker who does not wish to cause harm, though the decision maker may escape his/her (earthly) judgement through metaphorical death or merger.

The key result of the paper is that no matter how ethical one is, it is always possible to become trapped into knowingly committing an evil action which causes harm to people. This occurs if the bad news that the activity is harmful is received by the decision maker late enough – that is beyond their *evil time horizon*.

This mechanism can explain why a company run by an ethically minded CEO might nonetheless choose to hide evidence they discover that their product is addictive – perhaps such as the case of Oxycontin and the Sacklers, or the case of cigarettes causing cancer. The mechanism can also explain why a finance executive who did not wish to cause harm might nonetheless double down on cheating investment strategies to seek to avoid their judgement – perhaps such as the case of Bernie Madoff or Jérôme Kerviel.

Our analysis yields a number of policy implications. Better ethics among the population of decision makers creates benefits through multiple channels: better ethics lengthens the evil time horizon making honesty if bad-news appears more likely; better ethics increases the welfare created in the worst-case scenario of the new innovation being harmful; and the expected harm caused by innovators with better ethics is lower. Therefore selection of decision makers in favour of those with ethics would be desirable to a social planner. There are multiple ways in which ethics among decision makers can be improved. We discussed professional training programs, use of the fit-and-proper regime, and internal ethics training. Ethics are improved the more that decision makers internalise the experiences of

parties who would suffer from new innovations which transpired to be harmful. This would be an argument in support of greater exposure to the lived experience of those potentially affected by, for example, regular required volunteering at food pantries, food banks or even rehabilitation clinics.

Our second policy implication is that fines and detection are not substitutes. Better detection increases welfare as it increases the incentive for decision makers to confess the arrival of bad news without waiting to be forced by a public signal. But increasing fines will lower welfare if decision makers are ethical enough. Very large fines on firms can cause ex post damage to whole industries.[33] We show here that large fines can worsen decision maker behaviour *ex ante* also.

Though ethical decision makers provide many benefits our analysis shows that it is unethical decision makers who are most likely to win innovation contests. As innovation contests are widespread at present – to create usable AI and self driving cars for example – the dangers that unethical decision makers would win the contest and then conceal evidence that their invention is more harmful than thought seems a clear and present danger.

Finally the forces which give rise to the *evil time horizon* studied in this paper are deeply human ones. The tension at the heart of the model is perhaps best described by Virgil in The Aeneid:[34]

> The gates of hell are open night and day;
> Smooth the descent, and easy is the way:
> But to return, and view the cheerful skies,
> In this the task and mighty labor lies.

[33] See for example *British banks face an expensive motor finance crash*, Financial Times, November 8, 2024, link here.

[34] "Virgil, Aeneid 6.126–129 (Dryden's translation)" Cambridge, MA: Department of the Classics, Harvard University. I am grateful to David Storey for bringing this to my attention.

A Technical Proofs

Proof of Theorem 1. Let $F(t;t_0)$ be the value function of the agent after time t has elapsed since the receipt of the private bad-news signal at calendar time t_0. Therefore the calendar time since the start of the game is $t+t_0$. The agent must decide whether to cease the activity voluntarily and so admit that it has been harmful, or postpone the reckoning in the hope of being saved. The Bellman equation of the agent is

$$F(t;t_0) = \max\left[\underbrace{-\phi(t+t_0)}_{(A)}, \underbrace{\begin{array}{l}(\pi-\alpha v)dt - \phi(t+t_0)cdt + sdt\cdot 0\\ +\frac{1}{1+rdt}(1-sdt)(1-cdt)F(t+dt;t_0)\end{array}}_{(B)}\right]$$

Note that (A) denotes the termination value from stopping, while (B) captures the value of continuation. Let us consider the continuation region. Expanding out we establish that the value function satisfies:

$$\begin{array}{rl} F(t;t_0)rdt = & (\pi-\alpha v)dt + (F(t+dt;t_0) - F(t;t_0))\\ & -(s+c)F(t+dt;t_0)dt - \phi(t+t_0)cdt + o(dt)\end{array}$$

Which simplifies in the limit of $dt \searrow 0$ to yield the ode:

$$(r+s+c)F - \dot{F} = \pi - \alpha v - c\phi(t+t_0) \tag{11}$$

The homogeneous part of equation (11) has solution $\mathcal{A}e^{(r+s+c)t}$ for some constant $\mathcal{A}$. By inspection a particular solution exists which is linear in t and so we can establish that the general solution to (11) is:

$$F(t;t_0) = \mathcal{A}e^{(r+s+c)t} + \frac{1}{(r+s+c)}\left[\pi - \alpha v - c\phi(t+t_0) - \frac{c\phi}{r+s+c}\right]. \tag{12}$$

Let us define $\mathcal{F}(t;\tau,t_0)$ to be the value function of an agent time t after the

arrival of the private signal at t_0 who has the intention of waiting until τ has elapsed after the private signal was received at which point she will confess and accept her fine. The confession will therefore be made, unless saved or caught before, at calendar time $t_0 + \tau$. It follows that

$$F(t; t_0) = \sup_{\tau \geq t} \mathcal{F}(t; \tau, t_0) \tag{13}$$

$$\mathcal{F}(\tau; \tau, t_0) = -\phi(\tau + t_0) \tag{14}$$

Equation (13) notes that the optimal strategy of the agent at any point is determined by identifying the best time at which to confess. Equation (14) establishes the termination payoff (fine) that is payable if the confession is made τ units of time after the private signal was received (which was at calendar time t_0).

Using the general form for the value function in the continuation region, given by (12), we have that $\mathcal{F}(t; \tau, t_0)$ is of the form (12) with the constant $\mathcal{A}$ determined for each target confession delay τ. Denoting that $\mathcal{A}(\tau)$, it can be found using (14) to yield:

$$\mathcal{A}(\tau) = -\frac{1}{r+s+c} e^{-(r+s+c)\tau} \left[\pi - \alpha v + (r+s)\phi(t_0 + \tau) - \frac{c\phi}{r+s+c} \right]. \tag{15}$$

At the point that the private signal is received (so $t = 0$) the agent's value function is

$$\begin{aligned} \mathcal{F}(0; \tau, t_0) = &- \frac{1}{r+s+c} e^{-(r+s+c)\tau} \left[\pi - \alpha v + (r+s)\phi(t_0 + \tau) - \frac{c\phi}{r+s+c} \right] \\ &+ \underbrace{\frac{1}{r+s+c} \left[\pi - \alpha v - c\phi t_0 - \frac{c\phi}{r+s+c} \right]}_{(\dagger)} \end{aligned} \tag{16}$$

Where (16) is established by using the functional form (12) with the constant $\mathcal{A}(\tau)$ given by (15) and setting $t = 0$ to denote the moment the private bad news signal arrives.

The structure of the proof is to identify the delay τ which maximises $\mathcal{F}(0; \tau, t_0)$

and so yields the optimal strategy (13). We will establish this in two claims. To state these claims define the variable

$$G := \pi - \alpha v + (r+s)\phi t_0 - \frac{c\phi}{r+s+c} \tag{17}$$

We first claim that if $G \geq 0$ then it is optimal for the agent to wait indefinitely. To see this note that (using $(\dagger)$ defined in (16)) we have:

$$\begin{aligned} G \geq 0 \Rightarrow \mathcal{F}(0;\tau,t) &= (\dagger) - \frac{1}{r+s+c} e^{-(r+s+c)\tau}[G + \tau\phi(r+s)] \\ &\leq (\dagger)\ \forall \tau \geq 0, \end{aligned}$$

and the upper bound is achieved as $\tau \to \infty$. It follows that when the private information arrives (so $t = 0$) the optimal delay to target is ∞, i.e. the agent optimally never confesses and waits indefinitely.

The second claim is to show that if $G < 0$ then when the private bad-news signal is received (so $t = 0$) the optimal delay is 0, i.e. the agent confesses immediately. We have that

$$\begin{aligned} \mathcal{F}(0;\tau,t_0) &= (\dagger) - G \cdot \frac{1}{r+s+c} e^{-(r+s+c)\tau} - \frac{\phi(r+s)}{r+s+c} \tau e^{-(r+s+c)\tau} \\ &\leq (\dagger) - G \cdot \frac{1}{r+s+c} e^{-(r+s+c)\tau} \\ &\leq (\dagger) - G \cdot \frac{1}{r+s+c} \end{aligned}$$

And the supremum is achieved at $\tau = 0$. Therefore stopping immediately is optimal.

Therefore we have established that the agent will not confess to the action being harmful, and will continue with it indefinitely iff $G \geq 0$. This rearranges to yield that continuing with the harmful action, even though the agent knows it is

harmful is optimal iff:

$$t_0 \geq \frac{1}{\phi(r+s)}\left[\frac{c\phi}{r+s+c} + \alpha v - \pi\right]$$

Using this expression to define $t_{0*}(\alpha)$ as in (2) delivers the result. ■

Proof of Proposition 2. Differentiation of the evil time horizon, $t_{0*}(\alpha)$ given in (2). Note that $\alpha v > \pi \Rightarrow t_{0*} > 0$, which delivers the strict inequality in the second result. ■

Proof of Proposition 5. By hypothesis the activity is harmful. It therefore follows that a private bad-news signal will be received by the decision maker at some finite time, t_0, governed by the Poisson process with arrival rate b. If the signal arrives at time $t_0 < t_{0*}$ then by Theorem 1 the activity will cease and so the welfare from this point on will be 0. Capture this as $W_0^{stop} = 0$.

Consider now the case in which the private signal comes too late for the agent to be willing to confess: $t_0 > t_{0*}$. Define the welfare from continuing in this case as $W_0^{hide}(t)$ where t is the time elapsed after the private signal has arrived (at calendar time t_0). The evolution of welfare in this case is given by

$$W_0^{hide}(t) = (\pi - v)dt + \frac{1}{1+rdt}(1-sdt)(1-cdt)W_0^{hide}(t+dt) \tag{18}$$

Note that for welfare all the harm v is relevant to the welfare calculation, and that if the public signal arrives then the activity can be forced to stop. We also make the assumption that if the agent is 'saved' through death or a merger then the private signal is revealed and the activity therefore has to stop.

Expanding (18) and taking the limit as $dt \to 0$ yields the ode for welfare of:

$$(r+s+c)W_0^{hide} - \dot{W}_0^{hide} = \pi - v$$

The arrival rates of the public signal, and of salvation for the agent, are time independent. The flow of harm over profit is also time independent. It follows that

welfare is Markov and so $\dot{W}^{hide} = 0$ which yields that

$$W_0^{hide} \equiv -\frac{v-\pi}{r+s+c}. \tag{19}$$

We have therefore solved explicitly for the welfare at the point that the private signal arrives, conditional on the calendar time $t_0 \geq t_{0*}$.

We now need to discount these welfare values back to when the decision maker begins the activity, that is calendar time 0. Let us denote the welfare function before the private signal has been received $\{W^{hide}(t), W^{stop}(t)\}$ depending on whether the private signal will be received after or before t_{0*}. Note that t is now being used to denote time elapsed since the beginning of the game; that is calendar time. The time 0 value of welfare depends upon the possibility of the agent dying before the private signal can be received and also on the discount rate. The equation of evolution is given by

$$W^i(t) = (\pi - v)dt + \frac{1}{1+rdt}(1-sdt)W^i(t+dt) \quad \text{for } i \in \{hide, stop\}, \tag{20}$$

with boundary conditions

$$W^{hide}(t_0) = W_0^{hide} \text{ and } W^{stop}(t_0) = W_0^{stop} = 0. \tag{21}$$

Expanding (20) and taking the limit as $dt \to 0$ yields the differential equation:

$$(r+s)W^i - \dot{W}^i = \pi - v$$

And this has general solution:

$$W^i(t) = -\frac{v-\pi}{r+s} + A^i e^{(r+s)t}. \tag{22}$$

We can now solve for the npv of Welfare conditional on whether the agent will be confessing or hiding the news that the activity is harmful.

Consider first the case that $t_0 < t_{0*}$ so that the agent will confess and stop the

activity when the private signal of harm is received, at time t_0. Hence $W^{stop}(t_0) = 0$ as noted in (21). Substituting this boundary condition into (22) yields the constant of integration as

$$A^{stop} = \frac{v - \pi}{r + s} e^{-(r+s)t_0}$$

Therefore the welfare contribution at the beginning of the game from the setting in which $t_0 < t_{0*}$ is

$$W^{stop}(0) = -\left(\frac{v - \pi}{r + s}\right)\left[1 - e^{-(r+s)t_0}\right]. \tag{23}$$

Now consider the setting in which $t_0 > t_{0*}$ so that the agent will not confess when she gets the bad news signal. At time t_0 the welfare value will be given by W_0^{hide} established in (19). Therefore using this as the boundary condition in (22) we establish that

$$A^{hide} = \frac{(v - \pi)c}{(r + s)(r + s + c)} e^{-(r+s)t_0}$$

Therefore the welfare contribution at the beginning of the game in the setting in which the agent will not confess is, again using (22):

$$W^{hide}(0) = -\left(\frac{v - \pi}{r + s}\right)\left[1 - \frac{c}{r + s + c} e^{-(r+s)t_0}\right] \tag{24}$$

Turn now to the arrival times of the private bad news signal. The private bad-news signal arrives as a Poisson process with arrival rate b and so the arrival time of the bad news signal in the case, as assumed, that the practice is harmful is given by the exponential distribution: $f(t) = be^{-bt}$ (Cooper, 2005). So we can establish the overall expected welfare in the case when the activity is harmful:

$$W = \int_{t_0=0}^{t_{0*}} be^{-bt_0} W^{stop}(0) dt_0 + \int_{t_0=t_{0*}}^{\infty} be^{-bt_0} W^{hide}(0) dt_0$$

Using equations (23) and (24) we simplify to

$$W = -\left(\frac{v-\pi}{r+s}\right)\left[1 - \int_{t_0=0}^{t_{0*}} be^{-(r+s+b)t_0}dt_0 - \int_{t_0=t_{0*}}^{\infty} b\frac{c}{r+s+c}e^{-(r+s+b)t_0}dt_0\right] \tag{25}$$

Completing the integrations, this expression can simplified to:

$$W = -\left(\frac{v-\pi}{r+s}\right)\left[\frac{r+s}{r+s+b} + b\frac{r+s}{r+s+c}\left[\frac{e^{-(r+s+b)t_{0*}}}{r+s+b}\right]\right],$$

which simplifies to give the result. ■

Proof of Theorem 6. Denote the probability density function of ethics as h. Taking the expectation over ethics α we have:

$$\begin{aligned} E(W) =& \int_{\alpha=0}^{\infty} Wh(\alpha)d\alpha \\ =& W(1-H(0)) + \int_{\alpha=0}^{\infty} \frac{\partial W}{\partial \alpha}(1-H(\alpha))d\alpha, \end{aligned} \tag{26}$$

where we have used an integration by parts.

Next we use Proposition 5 to establish that

$$\frac{\partial W}{\partial \alpha} =_{\text{sign}} \frac{\partial t_{0*}}{\partial \alpha} \geq 0.$$

The final inequality following from Proposition 2. An immediate implication of substituting (4) into (26) is:

$$E_{H^{ethical}}(W) > E_{H^{unethical}}(W),$$

proving the result. ■

Proof of Theorem 7. Using Proposition 5 we establish that

$$\frac{\partial W}{\partial \phi} =_{\text{sign}} \frac{\partial t_{0*}}{\partial \phi}\begin{cases} < 0 & \text{if } \alpha v > \pi \\ \geq 0 & \text{if } \alpha v < \pi \end{cases},$$

where the second equality comes from Proposition 2. This yields the first and second result.

If $t_{0*} > 0$ then we can substitute the explicit value of (2) into Proposition 5 and so express welfare as:

$$W = -\mathcal{B}_0 \left[1 + \mathcal{B}_1 \cdot e^{\pi\psi} \cdot e^{-\alpha v \psi}\right]$$

where ψ is given in (6) and

$$\begin{aligned}\mathcal{B}_0 :=& \left(\frac{v-\pi}{r+s+b}\right)\\ \mathcal{B}_1 :=& \frac{b}{r+s+c} e^{-\frac{c(r+s+b)}{(r+s)(r+s+c)}},\end{aligned}$$

where both $\mathcal{B}_0, \mathcal{B}_1$ are both positive and independent of ϕ, using that $v > \pi$ by assumption. In the case that decision maker ethics are randomly distributed we calculate expected welfare as:

$$E_H(W) = -\mathcal{B}_0 \left[1 + \mathcal{B}_1 \cdot e^{\pi\psi} \cdot \int_{\alpha=0}^{\infty} e^{-\alpha v \psi} h(\alpha) d\alpha\right]$$

Differentiating expected welfare with respect to the fine rate then yields:

$$\begin{aligned}\frac{\partial}{\partial \phi} E_H(W) &= \frac{\partial}{\partial \psi} E_H(W) \underbrace{\frac{\partial \psi}{\partial \phi}}_{<0}\\ &= \underbrace{-\frac{\partial \psi}{\partial \phi}}_{>0} \mathcal{B}_0 \mathcal{B}_1 \frac{\partial}{\partial \psi} \left(e^{\pi\psi} \cdot M_h(-v\psi)\right)\\ &=_{\text{sign}} \left[\pi M_h(-v\psi) - v M_h'(-v\psi)\right].\end{aligned}$$

It follows that $\frac{\partial}{\partial \phi} E_H(W) > 0 \Leftrightarrow$ (5) holds. ■

Proof of Proposition 9. Using Proposition 5 we establish that

$$\begin{aligned}\frac{\partial W}{\partial c} &=_{\text{sign}} \frac{\partial}{\partial c}\left(\frac{-e^{-(r+s+b)t_{0*}}}{r+s+c}\right)\\ &=_{\text{sign}} \left[(r+s+c)(r+s+b)\underbrace{\frac{\partial t_{0*}}{\partial c}}_{>0}+1\right]e^{-(r+s+b)t_{0*}}\\ &> 0.\end{aligned}$$

The inequality in the middle line follows from Proposition 2. ■

Proof of Theorem 10. First note that V^{good} is independent of the decision maker's ethics as no harm is created if the activity is good.[35] But with probability q the decision maker anticipates that the activity will be harmful, and therefore lock-in is possible if the signal to this effect comes late enough. Theorem 10 therefore follows if we can show that

$$\partial V^{bad}/\partial \alpha < 0. \tag{27}$$

Suppose we are in the state of the world in which the activity is harmful. In this case V^{bad} will evolve until the private signal arrives at calendar time t_0. This time t_0 arrives according to a Poisson process and so the time of arrival is exponentially distributed. The value of $V^{bad}(t_0)$, with the argument capturing calendar time, will depend upon the whether t_0 is smaller or larger than the critical cut-off $t_{0*}(\alpha, \phi)$ which we established in Theorem 1. The agent's value function in either case is derived in the proof of Theorem 1.

We first derive the general form for $V^{bad}(t)$ as a function of calendar time since the beginning of the game. The game is not Markov as the boundary values are time dependent. Therefore, in advance of the private signal arriving, the equation

[35] One can establish that V^{good} evolves according to $V^{good} = \pi dt + \frac{1-sdt}{1+rdt}V^{good}$. The Markov nature of this setting then yields that

$$V^{good} = \frac{\pi}{r+s}.$$

of motion is:

$$V^{bad}(t) = (\pi - \alpha v)dt + \frac{1 - sdt}{1 + rdt} V^{bad}(t + dt)$$

This is analogous to the equation of motion of welfare established in (20). So analogously to (22) this has general solution

$$V^{bad}(t) = \frac{\pi - \alpha v}{r + s} + \mathcal{A} e^{(r+s)t} \tag{28}$$

for some constant $\mathcal{A}$.

Let us consider the case in which the private signal arrives too late for the agent to stop: $t_0 > t_{0*}(\alpha, \phi)$. It follows that at calendar time t_0 the agent's value function is given by $\lim_{\tau \to \infty}$ (16), which yields the boundary condition:

$$V^{bad}(t_0)|_{t_0 > t_{0*}} = \frac{1}{r + s + c} \left[\pi - \alpha v - c\phi t_0 - \frac{c\phi}{r + s + c} \right]$$

Using this boundary condition in (28) yields the constant $\mathcal{A}$. Therefore we can establish the value the agent anticipates at the beginning of the game:

$$V^{bad}(0)|_{t_0 > t_{0*}} = \frac{\pi - \alpha v}{r + s} \left[1 - \frac{c}{r + s + c} e^{-(r+s)t_0} \right] - \frac{c\phi}{r + s + c} e^{-(r+s)t_0} \left[t_0 + \frac{1}{r + s + c} \right]. \tag{29}$$

Now turn to the case in which the private signal will arrive soon enough that the agent will stop: $t_0 < t_{0*}(\alpha, \phi)$. In this case the agent will confess and accept the fine and so her value will be $-\phi t_0$ when the private signal arrives. This boundary condition applied to (28) yields the constant of integration $\mathcal{A}$ in this case. Therefore the value the agent anticipates at the beginning of the game in this case is

$$V^{bad}(0)|_{t_0 < t_{0*}} = \frac{\pi - \alpha v}{r + s} \left[1 - e^{-(r+s)t_0} \right] - \phi t_0 e^{-(r+s)t_0} \tag{30}$$

The value of the agent at the beginning of the game, allowing for the exponential

distribution of the private signal arrival time is therefore:

$$V^{bad} = \int_{t_0=0}^{t_{0*}} be^{-bt_0} V^{bad}(0)|_{t_0<t_{0*}} dt_0 + \int_{t_0=t_{0*}}^{\infty} be^{-bt_0} V^{bad}(0)|_{t_0>t_{0*}} dt_0$$

Then substituting in using (29) and (30) we have

$$\begin{aligned} V^{bad} =& \frac{\pi - \alpha v}{r+s} \left[1 - b \int_{t_0=0}^{t_{0*}} e^{-(r+s+b)t_0} dt_0 - b \frac{c}{r+s+c} \int_{t_0=t_{0*}}^{\infty} e^{-(r+s+b)t_0} dt_0 \right] \\ & - \phi b \left[\int_{t_0=0}^{t_{0*}} t_0 e^{-(r+s+b)t_0} dt_0 + \frac{c}{r+s+c} \int_{t_0=t_{0*}}^{\infty} e^{-(r+s+b)t_0} \left[t_0 + \frac{1}{r+s+c} \right] dt_0 \right]. \end{aligned}$$

We can now calculate $\partial V^{bad}/\partial \alpha$ and derive:

$$\begin{aligned} \frac{\partial V^{bad}}{\partial \alpha} =& - \left(\frac{v}{r+s+b} \right) \left[\frac{b}{r+s} + \frac{b}{r+s+c} e^{-(r+s+b)t_{0*}} \right] \\ & - \frac{\partial t_{0*}}{\partial \alpha} b e^{-(r+s+b)t_{0*}} \underbrace{\left(\frac{\pi - \alpha v}{r+s+c} + \frac{\phi}{r+s+c} \left((r+s)t_{0*} - \frac{c}{r+s+c} \right) \right)}_{\mathcal{D}} \end{aligned}$$

Now observe that $\partial t_{0*}/\partial \alpha \geq 0$ from Theorem 2, and from the definition of t_{0*} in (2) we have

$$\pi - \alpha v + \phi(r+s)t_{0*} - \frac{c\phi}{r+s+c} \geq 0,$$

which implies that $\mathcal{D} \geq 0$. Equation (27) now follows proving the result. ■

Proof of Proposition 11. The decision maker holding a project with ethics-harm pair (α, q) will launch the project if $qV^{bad}(\alpha) + (1-q)V^{good} \leq \kappa$. Define the worst project which would be launched by a type α decision maker as $\bar{q}(\alpha)$:

$$\bar{q}(\alpha)V^{bad}(\alpha) + (1 - \bar{q}(\alpha))V^{good} = \kappa$$

Note that some entry requires $\kappa < V^{good}$. Note also that

$$\frac{\partial V^{bad}(\alpha)}{\partial \alpha} < 0 \quad \text{(condition (27))} \quad \Rightarrow \frac{\partial \bar{q}(\alpha)}{\partial \alpha} < 0.$$

Now note that $\mathcal{E}_q(\alpha)$ is the conditional expectation of the probability of harm q conditional on $q < \bar{q}(\alpha)$. We can use an integration by parts to write this as:

$$\begin{aligned}\mathcal{E}_q(\alpha) &= E_J(q|q < \bar{q}(\alpha)) \\ &= \frac{\int_{q=0}^{\bar{q}(\alpha)} j(q)qdq}{J(\bar{q}(\alpha))} = \bar{q}(\alpha) - \int_{q=0}^{\bar{q}(\alpha)} \frac{J(q)}{J(\bar{q}(\alpha))}dq \\ &= \int_{q=0}^{\bar{q}(\alpha)} 1 - \frac{J(q)}{J(\bar{q}(\alpha))}dq\end{aligned}$$

And now note that

$$\frac{\partial}{\partial \bar{q}}\mathcal{E}_q(\alpha) = \int_{q=0}^{\bar{q}} \frac{J(q)}{J(\bar{q})^2} j(\bar{q})dq > 0$$

It therefore follows that $d\mathcal{E}_q(\alpha)/d\alpha < 0$ as required. ■

References

Abeler, J., D. Nosenzo, and C. Raymond (2019). Preferences for truth-telling. *Econometrica*.

Akerlof, G. A. and W. T. Dickens (1982). The economic consequences of cognitive dissonance. *The American economic review 72*(3), 307–319.

Arrow, K. (1973). Some ordinalist-utilitarian notes on Rawls's theory of justice. *Journal of Philosophy 70*(9), 245–263.

Bazerman, M. H., A. E. Tenbrunsel, and K. Wade-Benzoni (1998). Negotiating with yourself and losing: Making decisions with competing internal preferences. *Academy of management review 23*(2), 225–241.

Becker, G. S. (1968). Crime and punishment: An economic approach. In *The economic dimensions of crime*, pp. 13–68. Springer.

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological review 74*(3), 183.

Bénabou, R. and J. Tirole (2006). Incentives and prosocial behavior. *American Economic Review 96*(5), 1652–1678.

Bénabou, R. and J. Tirole (2011). Identity, morals, and taboos: Beliefs as assets. *Quarterly Journal of Economics 126*(2), 805–855.

Benmelech, E., E. Kandel, and P. Veronesi (2010). Stock-based compensation and ceo (dis) incentives. *The Quarterly Journal of Economics 125*(4), 1769–1820.

Bicchieri, C., E. Dimant, and S. Sonderegger (2023). It's not a lie if you believe the norm does not apply: Conditional norm-following and belief distortion. *Games and Economic Behavior 138*, 321–354.

Cooper, J. C. (2005). The poisson and exponential distributions. *Mathematical Spectrum 37*(3), 123–125.

Cressey, D. (1953). Other people's money; a study of the social psychology of embezzlement. *Patterson Smith*.

Daughety, A. F. and J. F. Reinganum (2005). Secrecy and safety. *American Economic Review 95*(4), 1074–1091.

Dixit, A. K. and R. S. Pindyck (1994). *Investment under uncertainty*. Princeton university press.

Doll, R. and A. B. Hill (1950). Smoking and carcinoma of the lung. *British medical journal 2*(4682), 739.

Dufwenberg, M. and M. A. Dufwenberg (2018). Lies in disguise–a theoretical analysis of cheating. *Journal of Economic Theory 175*, 248–264.

Dupont, Q. and J. M. Karpoff (2020). The trust triangle: Laws, reputation, and culture in empirical finance research. *Journal of Business Ethics 163*(2), 217–238.

Easley, D. and M. O'Hara (2023). Financial market ethics. *The Review of Financial Studies 36*(2), 534–570.

Ewing, J. (2017). *Faster, higher, farther: The inside story of the Volkswagen scandal*. Random House.

Fehr, E. and K. M. Schmidt (1999). A theory of fairness, competition, and cooperation. *The quarterly journal of economics 114*(3), 817–868.

Gennaioli, N., R. La Porta, F. Lopez-de Silanes, and A. Shleifer (2022). Trust and insurance contracts. *The Review of Financial Studies 35*(12), 5287–5333.

Gilbert, R. J. and D. M. G. Newbery (1982). Preemptive patenting and the persistence of monopoly. *American Economic Review 72*, 514–526.

Gneezy, U., A. Kajackaite, and J. Sobel (2018). Lying aversion and the size of the lie. *American Economic Review 108*(2), 419–453.

Golman, R., D. Hagmann, and G. Loewenstein (2017). Information avoidance. *Journal of economic literature 55*(1), 96–135.

Greene, J. and J. Haidt (2002). How (and where) does moral judgment work? *Trends in cognitive sciences 6*(12), 517–523.

Gul, F. and W. Pesendorfer (2001). Temptation and self-control. *Econometrica 69*(6), 1403–1435.

Haidt, J. (2007). The new synthesis in moral psychology. *science 316*(5827), 998–1002.

Hausman, D. and M. McPherson (1993). Taking ethics seriously: Economics and contemporary moral philosophy. *Journal of Economic Literature 31*(2), 671–731.

Hobbes, T. and M. Missner (2016). *Thomas Hobbes: Leviathan (Longman library of primary sources in philosophy)*. Routledge.

Hult, T. (1929). When ceos engage directly with customers. *Harvard Business Review*.

Kajackaite, A. and U. Gneezy (2017). Incentives and cheating. *Games and Economic Behavior 102*, 433–444.

Keefe, P. R. (2021). *Empire of pain: The secret history of the Sackler dynasty*. Anchor.

Leary, M. R. (2004, 08). Religion and Morality. In *The Curse of the Self: Self-Awareness, Egotism, and the Quality of Human Life*. Oxford University Press.

Leiper, R. (2017). Fitness and propriety. In *Conduct and Pay in the Financial Services Industry*, pp. 35–42. Informa Law from Routledge.

Lowell, J. (2012). Managers and moral dissonance: Self justification as a big threat to ethical management? *Journal of Business Ethics 105*(1), 17–25.

Malmendier, U. and G. Tate (2005). Ceo overconfidence and corporate investment. *The journal of finance 60*(6), 2661–2700.

Patel, K. (2014). *Setting Standards: Professional Bodies and the Financial Services Sec-*

tor. Centre for the Study of Financial Innovation (CSFI), London.
Schuchter, A. and M. Levi (2016). The fraud triangle revisited. *Security Journal 29*, 107–121.
Shu, L. L., F. Gino, and M. H. Bazerman (2011). Dishonest deed, clear conscience: When cheating leads to moral disengagement and motivated forgetting. *Personality and social psychology bulletin 37*(3), 330–349.
Sobel, J. (1985). A theory of credibility. *The Review of Economic Studies 52*(4), 557–573.
Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of economic behavior & organization 1*(1), 39–60.
Thanassoulis, J. (2023). Competition and misconduct. *The Journal of Finance 78*(4), 2277–2327.
Tirole, J. (1996). A theory of collective reputations (with applications to the persistence of corruption and to firm quality). *The Review of Economic Studies 63*(1), 1–22.
Vickers, J. S. (2011). *Independent Commission on Banking final report: recommendations*. The Stationery Office.
Vitell, S. J., M. Keith, and M. Mathur (2011). Antecedents to the justification of norm violating behavior among business practitioners. *Journal of Business ethics 101*, 163–173.
Wallace, J. and B. Thurman (2018). *Quantifying kindness, public engagement and place: experiences of people in the UK and Ireland*. Carnegie UK Trust.
Youngs, D. E., M. A. Yaneva, and D. V. Canter (2023). Development of a measure of kindness. *Current Psychology 42*(7), 5428–5440.