The Problem of Business Evil[*]

John Thanassoulis [†]

August 22, 2025

Abstract

We explain why ethical decision makers who care about others through other-regarding preferences can nevertheless be the cause of corporate scandals, such as *Barings*, *Société Générale* (both rogue trading), *Purdue Pharma* (opioid addiction), *WorldCom*, and *Bernie Madoff* (accounting/financial fraud). We show that all decision makers, however ethical, have an *in-too-deep threshold*. If private knowledge that the innovation they have introduced is harmful arrives after a threshold time then the decision maker will knowingly choose to continue to harm others – until publicly caught. We show the result is robust to patents and to limited liability. We show fines can make concealment of evil worse. And we show that ethical agents are more likely to lose innovation races, but punitive punishments after public revelation can mitigate the problem.

Keywords: Misconduct, corruption, ethics, fraud, fines

JEL codes: G41, D91

[*] My thanks to audiences at the keynote address for the Hellenic Finance and Accounting Association conference 2024, OLIGO 2025, CRETE 2025, 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

The *Problem of Evil*, studied for centuries, seeks to explain how a God can exist given evil exists. This paper seeks to explain how good people can exist given business evil exists. It is tempting to explain away every corporate scandal as an inevitable outcome of a key business executive being wicked. This paper explains why 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 on 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, which this paper models as having other-regarding preferences (Cooper and Kagel, 2016), can find themselves knowingly perpetuating behaviour they know is bad and harmful. The model presented will offer an explanation as to how even good people could knowingly commit bad business acts such as:

1. **Nick Leeson** and Barings Bank rogue trading. Leeson's role was to execute client trades, making money on pricing differences between Singapore and Japan.[3] Leeson began to take illegitimate bets which initially performed well. When the trades started loosing money the losses could be hidden in an 'error' account he had access to. In 1995 the Kobe earthquake caused prices to diverge overnight such that the loss could not be hidden. Nick Leeson faced the potential of 84 years in jail. Leeson said his actions weren't due to greed, *"it was a fear of failure...I couldn't tell anybody about it."* (Newlands, 2025).

2. **Jérôme Kerviel** and Societe Generale financial investments fraud. Kerviel 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).[4] However the bets then started losing money and eventually this became so substantial that Kerviel was discovered. He was sentenced in court to pay €4.9billion. Yet all of

[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] Nick Leeson: the man who broke Barings Bank, MoneyWeek, 1 July, 2000

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

this was seemingly not motivated by greed: Forbes reports *"he apparently made no personal profit from his trades"*.[5]

3. **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[6]) – 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.[7] 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.

4. **WorldCom Scandal**:[8] 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."*

5. **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 gave significant sums to charity through his multi-million dollar foundation[9], and at trial he declared his repentance 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."*

To study these, and similar episodes, we define rigorously and solve the following problem. A decision maker undertakes an activity which he thinks is very likely benign or even beneficial. 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

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

[6] See Ohio Detox Centre for example (link).

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

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

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

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 (metaphorical) 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. After the private signal, a public signal will arrive, but exactly when is not known. 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 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 fine acts to deter the decision maker from revealing the arrival of bad news. The second force in our model is the effect of ethics. If the decision maker chooses to ignore the private bad news signal and continue the activity then he knows he is harming the clients. If the decision maker is ethical then this will reduce his 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. If the public bad news signal arrives with the decision maker still in post then he will be forced to stop and he 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 (a little) 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 the 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 always exists an *in-too-deep threshold* such that if the private bad news signal arrives after this time has elapsed since the commencement of the innovation, the decision maker will knowingly prefer to continue with the innovation, harming the clients and hiding the news that the innovation is harmful.

Continuing with a known harmful activity is painful for the decision maker – the harm knowingly inflicted hurts the decision maker, and 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 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 link between the length of the in-too-deep threshold and the fine rate is surprising. If the decision maker places sufficient weight on the clients' well-being (i.e. is sufficiently ethical) then increasing fines *shortens* the in-too-deep threshold. 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 in-too-deep threshold. If the private bad news signal arrives right on the in-too-deep threshold then the ethical decision maker dislikes equally stopping the activity immediately and incurring the fine versus putting off a larger fine plus the ongoing pain which will be caused to clients from continuing. As carrying on is ethically painful for the decision maker, it follows that the immediate fine from confessing is larger than the discounted fine from continuing. If the fine rate rises therefore 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 in-too-deep threshold moves earlier.

This logic offers a new perspective as to how rogue trading (Barings, Societe Generale), accounting fraud (WorldCom), and Ponzi schemes (Madoff) begin.

If the business opportunity ends, e.g. a patent expiring, then the in-too-deep threshold 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 maker will prefer to conceal the knowledge and knowingly continue to harm through to the expiration of the patent. This offers a new perspective as to how an ethical decision maker would deliberately decide to continue with a known harmful drug through to the end of the patent – perhaps such as the case of OxyContin and the Sackler family.

We next explore the case in which the decision maker cannot be punished beyond some maximum level. This might be the case, for example, if prison sentences cannot be enforced, if the decision maker is shameless, and of limited financial means. We show the in-too-deep threshold applies for all except those who are the most ethical *and* also the poorest. For everyone except these few the in-too-deep-threshold exists: accepting the maximum bearable fine by confessing is more painful than knowingly choosing to cause harm plus the (discounted) maximum bearable fine from continuing the harm.

In an innovation race we show ethical agents are disadvantaged and more likely to lose to unethical competitors. This might be surprising as unethical agents anticipate they are unlikely to admit to bad news and so risk the largest fines. Sadly we also show that the expected harm from new innovations grows the less ethical the innovators are. These two results have troubling implications for innovation races such as those now being conducted in AI and for self-driving cars.

We also study the benefits of punishment programs in which the fine for bad actions differs depending upon whether the agent stops the activity before or after its harm becomes public knowledge. We demonstrate that the in-too-deep threshold applies for moderate punishment regimes, but if the extra punishment after involuntary cessation is large enough the in-too-deep threshold can be bounded in time.

Contribution to the 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 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 find themselves choosing to commit misconduct. Business Ethics research has a long history in characterising the ingredients which they consider 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 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 he is committing the fraud. The decision maker we study in this paper has all of these ingredients.[11]

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

[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.

2 The Model

A decision maker has decided to undertake a possibly, but he hopes unlikely to be, harmful activity. Examples are selling a drug after proper 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 favour.

At the point of commencing the decision maker realises the activity might be harmful and places a probability q that this is the case. With probability $1 - q$ the decision maker believes the activity is not harmful. We assume that the decision maker decides to start the activity at $t = 0$. 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. Hence if harmful the activity is welfare destructive. 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), Thanassoulis (2023)). Thus α measures the weight given to third party utility in the decision maker's utility function. 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's game is terminated with probability sdt. We think of this as the agent being is 'saved' which could occur by merger, by death or some other *force majeur*. 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 bdt 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 bad-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'. 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. In the benchmark model there is no difference between the fine rate received after voluntary cessation versus involuntary cessation. If there were such a difference in fines then the decision maker could front-run the public signal by admitting liability just before the public authorities were able to go to court and enforce against the public signal.

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, and 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 acting on the injured parties' behalf.[14] In the benchmark model 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]). We exlore the effect of limited liability in an

[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.

[14] For more discussion see 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.

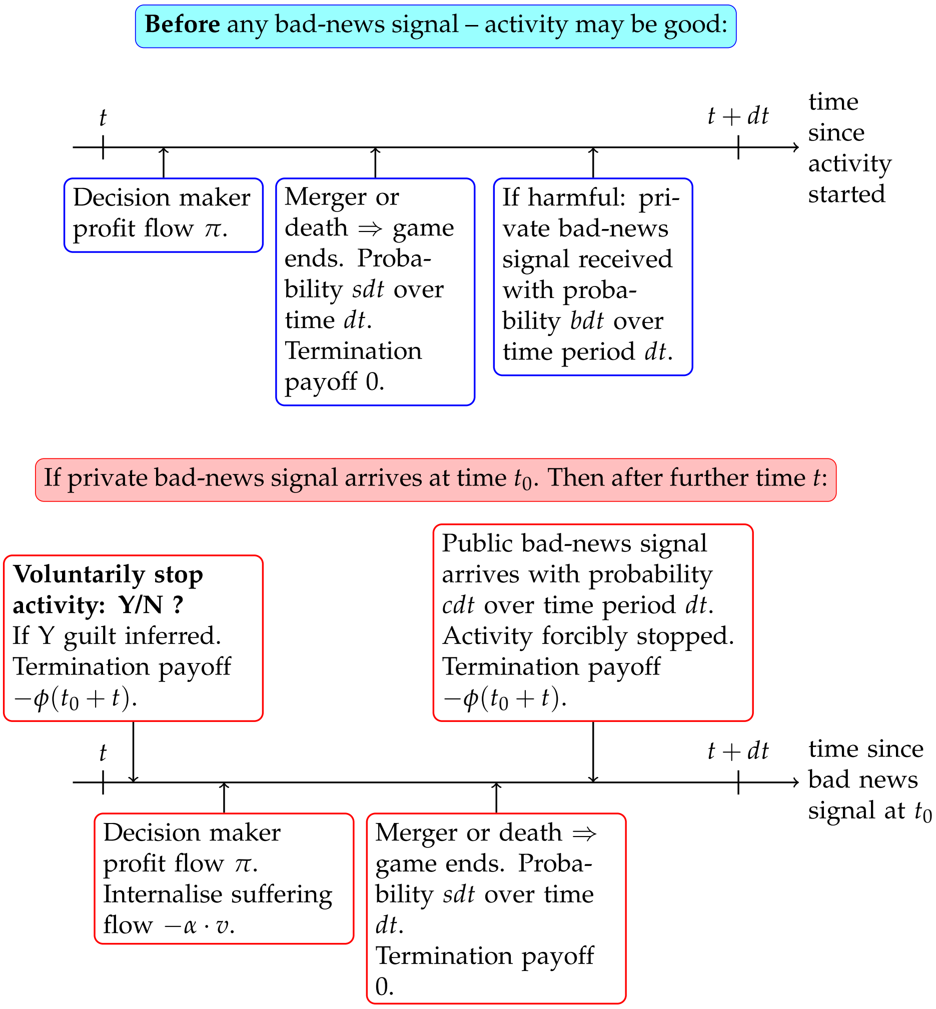
Before any bad-news signal – activity may be good:
t
$t + dt$
time since activity started
Decision maker profit flow π.
Merger or death ⇒ game ends. Probability sdt over time dt. Termination payoff 0.
If harmful: private bad-news signal received with probability bdt over time period dt.
If private bad-news signal arrives at time t_0. Then after further time t:
Voluntarily stop activity: Y/N ? If Y guilt inferred. Termination payoff $-\phi(t_0 + t)$.
Public bad-news signal arrives with probability cdt over time period dt. Activity forcibly stopped. Termination payoff $-\phi(t_0 + t)$.
t
$t + dt$
time since bad news signal at t_0
Decision maker profit flow π. Internalise suffering flow $-\alpha \cdot v$.
Merger or death ⇒ game ends. Probability sdt over time dt. Termination payoff 0.

Figure 1: Model timeline

extension.

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).

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

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

to the setting in which the business opportunity has a defined end date, as in the case of a patent, in Section 5.

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 control. The results we will present are robust to any such embellishment of the model.[18]

3 The business evil problem

Suppose that a decision maker begins a new activity at calendar time $t = 0$. Our first result establishes that no matter how ethical the decision maker is, he can always find himself choosing to knowingly harm others:

Theorem 1 *No matter how ethical the decision maker is* ($\alpha \in \mathbb{R}_+$) *there exists an* in-too-deep threshold $t_{0*}(\alpha) \geq 0$ *such that if the decision maker discovers the activity is harmful after time* $t_{0*}(\alpha)$ *since business launch then the decision maker will choose to continue with the known harmful activity until he is either saved by merger or death or the authorities receive the public bad-news signal causing cessation to be enforced.*

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

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

[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.

other event relieving him 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 has not admitted that the action is harmful for a further t units of time is:

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

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 explain the intuition for the result before describing the proof steps, with the formal proof in the Appendix.

Theorem 1 establishes that no matter how ethical the decision maker is, that is no matter how intensely 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 he is undertaking is harmful.

The intuition for this is as follows. Continuing with the activity causes the decision maker pain as 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 he admits the activity is wrong. Theorem 1 shows that no matter how ethical the decision maker is, there is always a threshold time such that bad news arriving after this point will be suppressed. We call this point the in-too-deep threshold.

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[19]

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

Description of proof steps

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 function 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 (14)) captures 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, 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 (15). 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.

[19] Click here for the Madoff trial transcript.

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 *in-too-deep threshold.* 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 he is ethical, never confesses and carries on perpetuating the harmful activity.

4 Characteristics of the in-too-deep threshold

The proof of Theorem 1, and in particular at (16), establishes that the in-too-deep threshold is given by:

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

We now interrogate the characteristics of the in-too-deep threshold.

Proposition 2 *The in-too-deep threshold ($t_{0*}(\alpha)$) beyond which the ethical agent will knowingly continue to harm satisfies:*

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

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

3. *Better public detection (c ↑) causes the decision maker to have a larger in-too-deep threshold (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.

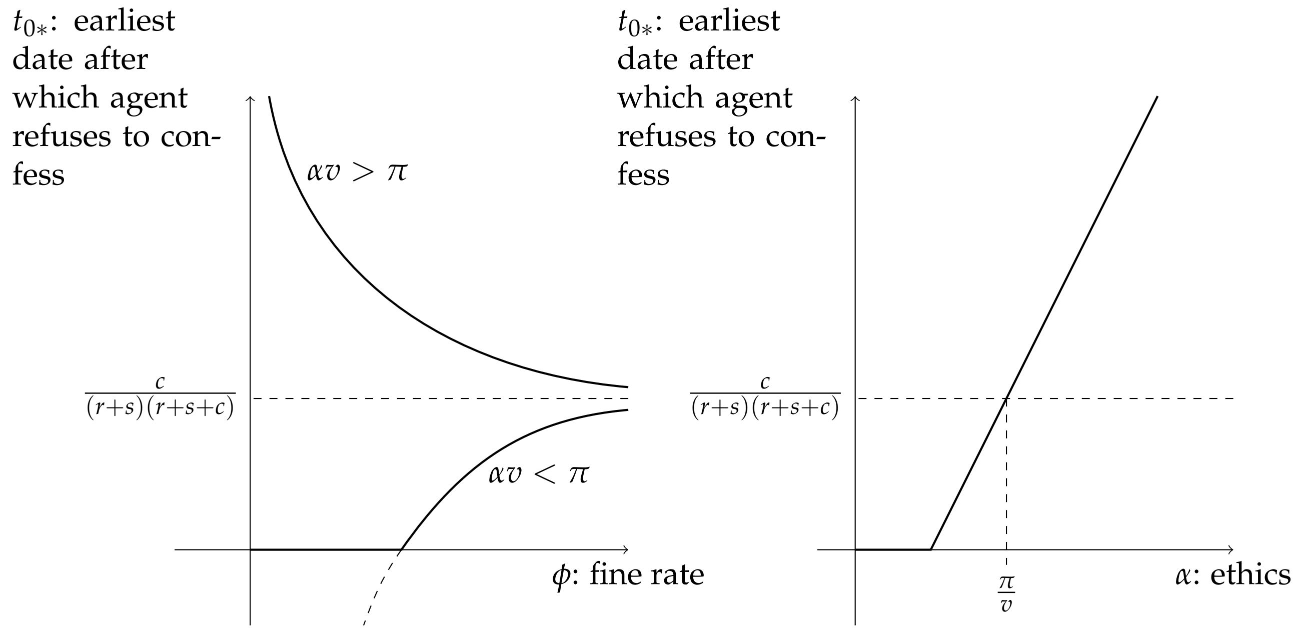

Figure 2: The last point at which agent is willing to confess: t_{0*}.

Notes: The in-too-deep threshold is the critical time after which the agent will not confess that the activity is harmful. The in-too-deep threshold 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 in-too-deep threshold 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 in-too-deep threshold 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 in-too-deep threshold moves out.

The intuition for the fact that increasing penalties brings forward the in-too-deep threshold and so brings forward the date of lock-into evil for the most ethical is more subtle. Recall that the in-too-deep threshold 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, versus the alternative of continuing with the practice despite the harm it causes. 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:

$$\left.\begin{array}{l}\text{Indifference}\\ \text{between}\\ \text{stopping and}\\ \text{continuing}\end{array}\right\} \Rightarrow G|_{t_{0*}} = 0 \Leftrightarrow \underbrace{\phi(r+s)t_{0*}}_{\substack{\text{fine from}\\ \text{confessing}\\ \text{immediately}}} = \underbrace{\phi\frac{c}{r+s+c} + \alpha v - \pi}_{\substack{\text{payoff rate}\\ \text{from continuing}}} \tag{3}$$

The time t_{0*} occurs when the payoffs from confession versus continuation are equal. If $\alpha v > \pi$ then the flow of utility from denial is itself negative (as the ethics cost exceeds the profit gain for every moment of denial). So the fine incurred from immediate confession must be greater than the fine npv component from indefinite waiting for the agent to be indifferent between confessing and not. This implies that the sensitivity to the fine rate of immediately confessing must be larger than the sensitivity to the fine rate from denial.[20] 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. It follows that the in-too-deep threshold, the lock-in moment, t_{0*}, moves forward in time as the fine rate ϕ increases. Hence fines are counter-productive if the agent is ethical.

An implication of Proposition 2 is therefore that whether penalties should be increased or decreased depends upon the ethics of the decision makers in the population. In unreported work we can establish that if ethics are randomly distributed in the population then whether Welfare is increased or otherwise by lowering punishments depends upon the moment generating function of the distribution of ethics in the population.

A further, I think surprising, result is the robustness of the in-too-deep threshold 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.

[20] Mathematically, using (3) above

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

Corollary 3 *Suppose that the decision maker does not internalise any of the profits made from sale then there remains a positive in-too-deep threshold 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 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.[21]

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 punishment or shame. But in the case of continuing, even though there is no off-setting 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 in-too-deep threshold.

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 in-too-deep threshold. 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 early 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 in-too-deep threshold 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

[21] See footnote 8.

dominate. In this case the decision maker would rather confess and stop the harmful activity.

5 The in-too-deep threshold applies to patents

In the model we have studied the business activity did not have a defined end date. The in-too-deep threshold 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 find that the in-too-deep threshold result is robust: it is always possible for an ethical decision maker to choose to knowingly harm his clients.

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 – though he can still be fined if evidence of harm comes to light. The rest of the model is unchanged: the decision maker discounts the future, he 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 up to $-\phi T$. And this fine applies if the agent is discovered even after the patent has ended.

Proposition 5 *However ethical a decision maker is a unique in-too-deep threshold $t_{0*}(\alpha) < T$, shorter than the length of the business opportunity, exists.*

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 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 in-too-deep threshold, 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 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.

6 Limited Liability

The in-too-deep threshold arises as the decision maker who finds out late that his actions are harmful faces a high penalty for admitting this, and a potentially even higher penalty for lying. In this section we establish that the result is generally robust, even if the decision maker enjoys a limit on how much he can be punished. Such limited liability would need prison to be unavailable, the agent to be shameless, and his wealth to be low enough.

Proposition 6 *Suppose that a decision maker cannot be fined more than W.*

1. *If*
$$\pi - \alpha v + (r+s)W > 0 \tag{4}$$
then there is a finite in-too-deep threshold, $t_{0*} \in [0, W/\phi)$. If the bad news comes before t_{0*} then the agent will confess that the new product is bad. But bad news after t_{0*} will be ignored and the decision maker will continue with the known harmful activity.

2. *If* (4) *does not hold then there is no finite in-too-deep threshold. The decision maker always confesses immediately on the receipt of bad news.*

We see that under limited liability for all except the most ethical *and also* the poorest, a finite *in-too-deep threshold* exists. Hence the message of this paper that there is a time after which ethical decision makers will knowingly prefer to harm third parties is robust.

We first give the intuition before we describe the method of proof.

We plot the result of Proposition 6 in Figure 3. To determine the intuition suppose that a private signal is received after so much time has elapsed since the decision maker started the activity that stopping now would entail a fine in excess of the maximum bearable: W. If the decision maker nevertheless confesses then he accepts the maximum fine he can bear: W. If however the decision maker decides to hide the bad news then he now knowingly does harm, and this creates a flow utility of $\pi - \alpha v$, this is the profit less the pain of knowingly harming. This flow utility is negative if the agent is ethical enough. In addition the public signal will arrive and the decision

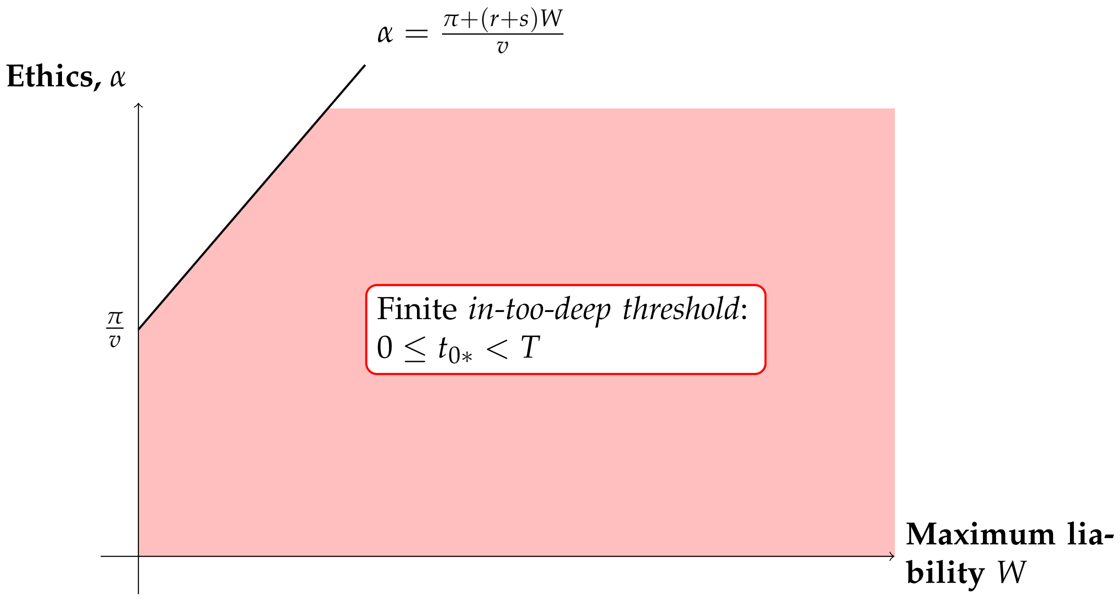

Figure 3: *In-too-deep threshold* with maximum liability
Notes: The region in which a finite in-too-deep threshold exists is bounded by (4) given in Proposition 6. The time at which the maximum liability becomes binding is $T = W/\phi$. A finite in-too-deep threshold exists for all except the poorest *and* most ethical. These agents always confess if the private bad news signal arrives.

maker will have to pay the maximum fine W in the future, unless the random exit event arrives. Allowing for discounting these two effects (exit and discounting) reduce the anticipated fine by $(r+s)W$. Combining we see why the decision maker has an in-too-deep threshold if condition (4) is satisfied. That is the in-too-deep threshold exists for all except the simultaneously poorest and most ethical.

The proof contains two distinct parts as the decision maker's problem differs depending on whether he has been in the market for so long that the fine would exceed the limited liability protection, or otherwise. The method of proof is to use backward induction and derive the optimal strategy if the limited liability constraint is binding. The proof then repeats the analysis if the private bad news signal arrives early enough that the fine following immediate confession is below the limited liability constraint.

As the decision maker cannot be fined more than W there is a point in time, we label $T = W/\phi$, such that after T the fine remains constant at the maximum level of W. The first part of the proof is to determine whether, if the private bad news signal should arrive at or after this time T, the decision maker would confess, or would hide the news. We derive the value function for a decision maker who received bad news at $t_0 \geq T$ and planned to wait for a further time τ to elapse before confessing. Optimising over the waiting time τ we derive that the optimal strategy hinges on the sign of (4). We establish that if (4) holds then the decision maker will prefer never to confess. If (4) doesn't hold then the decision maker will prefer to confess immediately.

The next step is then to use the insights above to derive the value of the decision maker at the point in which limited liability becomes binding (calendar time T). This terminal value, which we denote $F(T)$, takes on two values depending on whether (4) holds. We can now derive the value function of the decision maker if he receives the private bad news before the limited liability point, i.e. $t_0 < T$. This analysis parallels that of the benchmark model and so we immediately have the value function if the decision maker should wait for any time $\tau < T - t_0$ so that limited liability is not triggered. These values need to be compared to the value of waiting until exactly calendar time T.

The proof establishes that the value function is quasi-convex and so waiting for intermediate times is never optimal: either it is best to confess immediately or to wait until calendar time T and then proceed optimally, completing the proof.

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. More ethical agents will be more concerned about harmful innovations. However they will also stop a harmful innovation sooner if bad news comes to light, limiting the harm done and the fine imposed. It is not therefore clear whether the expected value of entry is declining in ethics, or not.

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 $V^{bad}(\alpha)$, and V^{good} if the activity is benign. The agent is willing to expend a value $\bar{V}$ to commence the project where:

$$\bar{V}(\alpha) := qV^{bad}(\alpha) + (1-q)V^{good}. \tag{5}$$

Proposition 7 *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 Proposition 7 is to establish the value which a decision maker can expect should the product they are launching turn out to be harmful. 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.

Proposition 7 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 Proposition 7 is related to the envelope theorem. If we were to increase an agent's level of ethics marginally then she would be less vulnerable to being locked into lying and be more willing to confess if and when a bad-news private signal is received. As the critical lock-in, the in-too-deep threshold 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.

The model predicts therefore that less ethical agents would win a contest to launch a 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 Proposition 7 and so generate an empirical hypothesis concerning the expected quality of products introduced as a function of decision maker ethics.

Proposition 8 *Suppose that decision maker ethics* (α) *and the probability of innovation harmfulness* (q) *are randomly and independently distributed. Suppose that entry costs are fixed at an amount: κ. Projects which are launched and run by more ethical decision makers are safer on average.*

Proposition 8 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 of Proposition 7 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 an integral argument establishes the result that conditional on entry, more ethical agents have products which are less likely to be harmful.

8 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,[22] the SEC is accused of demanding small fines,[23] and plea deals are a feature of US justice.[24]

Extra fine regimes have $\phi^{extra} > 0$ and so they increase the fine rate if cessation is forced. However such fines can be evaded if the decision maker can discover when a public signal is about to arrive, e.g. by monitoring relevant publications, whistleblowers, or the authorities themselves. Nonetheless, we also consider:

2. Define **punishment** as the case

$$0 < \phi^{extra} < \left(\frac{r+s}{c}\right)\phi \tag{6}$$

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

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

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

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

$$\phi^{extra} > \left(\frac{r+s}{c}\right)\phi \tag{7}$$

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:

Proposition 9 *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.*

1. *Theorem 1 applies in the* ***leniency****, and* ***punishment*** *regimes:*
   for any level of decision maker ethics, $\alpha \in \mathbb{R}_+$, there exists an in-too-deep threshold $0 \leq t_{0*}(\alpha) < \infty$ *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 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*

   $$\alpha v \geq \pi - \phi \tag{8}$$

   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 $\mathcal{T}_{0*}(\alpha)$.*

Proposition 9 is proved using the techniques we have established in Theorem 1 and in Proposition 5. 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 in-too-deep threshold 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 (8) 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 and the agent might be saved. 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 then the decision maker will be faced with an extra fine, so overall 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 (7):

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

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 Proposition 9.

If the authorities can increase the fine payable by the decision maker on the occasion that the authorities force cessation then the in-too-deep threshold 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.

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 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 (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 *in-too-deep threshold*.

The mechanism can explain why a finance executive who did not wish to cause harm might nonetheless double down on cheating investment strategies to seek to avoid judgement – perhaps such as the case of Nick Leeson, Bernie Madoff or Jérôme Kerviel. This mechanism can also 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.

We show that our result extends to patents. Here the in-too-deep threshold is brought forward so that decision makers would not reveal bad news just before the end of a patent. Limited liability leaves the result largely robust. Extensions of this work are possible: such as to allow for time varying probabilities of signals arriving, or to allow decision makers to seek to impede the revelation of public bad news.

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 in-too-deep threshold making honesty if bad-news appears more likely; 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. These include 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. How one might evidence this is an open question.

Our second policy implication is that increasing fines will make the problem of the in-too-deep problem worse if decision makers are ethical enough. Very large fines on firms can cause ex post damage to whole industries.[25] We show here that large fines can worsen decision maker behaviour *ex ante* also. However punishments which become punitive in the event that private information is not admitted to can mitigate the problem.

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 must be a concern.

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

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

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

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

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

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

$$(r+s+c)F - \dot{F} = \pi - \alpha v - c\phi(t+t_0) \tag{9}$$

The homogeneous part of equation (9) 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 (9) is:

$$F(t;t_0) = \mathcal{A}e^{(r+s+c)t} + \frac{1}{(r+s+c)}\left[\pi - \alpha v - c\phi(t+t_0) - \frac{c\phi}{r+s+c}\right]. \tag{10}$$

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

$$F(t;t_0) = \sup_{\tau \geq t} \mathcal{F}(t;\tau,t_0) \tag{11}$$

$$\mathcal{F}(\tau;\tau,t_0) = -\phi(\tau+t_0) \tag{12}$$

Equation (11) notes that the optimal strategy of the agent at any point is determined by identifying the best time at which to confess. Equation (12) establishes the termination payoff (fine) that is payable if the confession is made at calendar time $t_0 + \tau$, i.e. 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 (10), we have that $\mathcal{F}(t; \tau, t_0)$ is of the form (10) with the constant $\mathcal{A}$ determined for each target confession delay τ. Denoting that $\mathcal{A}(\tau)$, it can be found using (12) to yield:

$$\mathcal{A}(\tau) = -\frac{1}{r+s+c} e^{-(r+s+c)\tau} \left[\pi - \alpha v + (r+s)\phi(t_0 + \tau) - \frac{c\phi}{r+s+c} \right]. \tag{13}$$

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

$$\begin{aligned} \mathcal{F}(0; \tau, t_0) = &- \frac{1}{r+s+c} e^{-(r+s+c)\tau} \left[\pi - \alpha v + (r+s)\phi(t_0 + \tau) - \frac{c\phi}{r+s+c} \right] \\ &+ \underbrace{\frac{1}{r+s+c} \left[\pi - \alpha v - c\phi t_0 - \frac{c\phi}{r+s+c} \right]}_{(\dagger)} \end{aligned} \tag{14}$$

Where (14) is established by using the functional form (10) with the constant $\mathcal{A}(\tau)$ given by (13) 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 (11). We will establish this in two claims. To state these claims define the variable

$$G := \pi - \alpha v + (r+s)\phi t_0 - \frac{c\phi}{r+s+c} \tag{15}$$

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 (14)) 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

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

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:

$$t_0 \geq \frac{1}{\phi(r+s)}\left[\frac{c\phi}{r+s+c} + \alpha v - \pi\right] \tag{16}$$

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

Proof of Proposition 2. Differentiation of the in-too-deep threshold, $t_{0*}(\alpha)$ 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. The proof proceeds in a number of steps.

We first evaluate the value function at calendar time T, the end of the business opportunity, if the private bad news signal has been received then or earlier, $t_0 \leq T$. If the decision maker admits that the activity is harmful then she is fined: payoff $-\phi T$. If instead the decision maker hides the bad news and the activity ends then the value is Markov (history independent). Denote this value V and note that it satisfies:

$$V = -\phi T c dt + \frac{1}{1+rdt}(1-sdt)(1-cdt)V$$

There is no harm done or profit earned over each dt period of time. If not saved and not caught then the value function continues, appropriately discounted. In the limit of $dt \to 0$ this solves to yield

$$V = \frac{-c\phi T}{r+s+c}$$

It is immediate that $V > -\phi T$, the payoff from admitting. Hence we have shown that at calendar time T, the end of the business opportunity, all decision makers will deny having received any bad news. Hence the terminal value at this point is

$$F(T) = V = \frac{-c\phi T}{r+s+c} \tag{17}$$

The next step of the proof is to establish the payoff to an agent who receives a private bad news signal at time $t_0 < T$ and considers waiting for a duration $\tau \in [0, T - t_0)$ before admitting that the activity is harmful. This is the problem solved in Theorem 1 and the value function at calendar time t_0 (i.e. after time $t = 0$ has elapsed since the private bad news signal is received) is given by $F(0; \tau, t_0)$ given in (14).

The next step is to establish the value of waiting, upon receipt of private bad news, to the moment the business opportunity ends: $F(0; T - t_0, t_0)$. We have established that at the end point the terminal value will be given by $F(T)$ explicitly given in (17). The evolution of the value function in the continuation region is given by (10). So setting $F(T - t_0; t_0) = F(T)$ yields the constant in the solution of the ode, $\mathcal{A}(T - t_0)$:

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

and so substituting this into (10) and evaluating at the point the bad news signal arrives we have

$$F(0; T - t_0, t_0) = \frac{1}{r+s+c}\left[\pi - \alpha v - \frac{c\phi}{r+s+c}\right]\left(1 - e^{-(r+s+c)(T-t_0)}\right) - \frac{c\phi t_0}{r+s+c} \tag{18}$$

Now we observe that there is a jump up in the value function if the decision maker waits to time T and so optimally hides at that point versus admitting the bad news just before calendar time T:

$$\underbrace{\lim_{\tau \to T - t_0} F(0; \tau, t_0)}_{\text{Using (14)}} < \underbrace{F(0; T - t_0, t_0)}_{\text{Using (18)}}$$

This is clear by inspection.

The next step in the proof is to establish that $F(0; \tau, t_0)$, given in (14) is quasi-convex in τ. To do this we note that for $\tau \in (0, T - t_0)$:

$$\frac{\partial}{\partial \tau} F(0; \tau, t_0) = e^{-(r+s+c)\tau}\left[\pi - \alpha v + (r+s)\phi(t_0 + \tau) - \frac{c\phi}{r+s+c} - \frac{(r+s)\phi}{r+s+c}\right] \tag{19}$$

This has only one root in τ, call it $\tilde{\tau}$, and further $(r+s)\phi > 0 \Rightarrow \left.\frac{\partial F}{\partial \tau}\right|_{\tau < \tilde{\tau}} < 0 < \left.\frac{\partial F}{\partial \tau}\right|_{\tau > \tilde{\tau}}$. Hence quasi-convexity is established.

The above implies that if bad news is revealed then the optimal strategy for the decision maker is either to admit immediately, paying $-\phi t_0$, or to hide the bad news up to the end of the business opportunity at calendar time T and beyond. Using (18) the agent will admit immediately if and only if

$$-\phi t_0 > F(0; T - t_0, t_0)$$

$$\text{(using (18))} \quad -(r+s)\phi t_0 > \underbrace{\left[\pi - \alpha v - \frac{c\phi}{r+s+c}\right]}_{(\mathcal{C})} \left(1 - e^{-(r+s+c)(T-t_0)}\right) \qquad (20)$$

Now we note that if $(\mathcal{C}) \geq 0$ then (20) cannot be satisfied and so the decision maker will never admit to bad news: $t_{0*} = 0$.

Suppose that $(\mathcal{C}) < 0$. We apply the Intermediate Value Theorem noting that the left hand side of (20) is declining in t_0, while the right hand side is increasing. It can therefore be shown that there exists a unique value of t_0, labelled $t_0* \in (0, T)$, such that (20) is satisfied for all $t_0 < t_{0*}$. The in-too-deep threshold is defined implicitly as:

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

This completes the proof. ■

Proof of Proposition 6 . Let us denote by $T = W/\phi$ the point in time at which the limited liability of the decision maker kicks in.

We derive the equation giving the evolution of the value function in the continuation region when the decision maker is beyond his limited liability point of T. The continuation value $F(t; t_0)$ for time t after the private bad news arrived at $t_0 \geq T$ satisfies:

$$\begin{aligned} F(t; t_0) =& (\pi - \alpha v)dt - \phi T c dt + 0 \cdot s dt \\ &+ \frac{1}{1 + r dt}(1 - s dt)(1 - c dt)F(t + dt; t_0) \end{aligned}$$

Taking the limit as $dt \searrow 0$ generates the ode:

$$(r+s+c)F - \dot{F} = \pi - \alpha v - \phi T c$$

So evaluating the homogeneous solution and the particular solution we derive the general form for the continuation value after time T as:

$$F(t; t_0) = \mathcal{A} e^{(r+s+c)t} + \frac{\pi - \alpha v - \phi T c}{r+s+c}. \qquad (21)$$

We now derive the value function for a decision maker who is beyond his limited liability point and intends to wait for time τ after the private news arrives (at $t_0 > T$) before confessing. We denote this $\mathcal{F}(t; \tau, t_0)$. In the continuation region this function will take the form (21). We identify the constant, $\mathcal{A}(\tau)$ by using the boundary condition that the decision maker will confess at calendar time $t_0 + \tau$. This yields:

$$\mathcal{A}(\tau)e^{(r+s+c)\tau} + \frac{\pi - \alpha v - \phi Tc}{r+s+c} = -\phi T$$
$$\Rightarrow \mathcal{A}(\tau) = -\frac{1}{r+s+c}e^{-(r+s+c)\tau}[\pi - \alpha v + (r+s)\phi T]. \tag{22}$$

And so the value function of the decision maker at the moment the private bad news signal arrives after the limited liability point $(t_0 > T)$ and who intends to wait for time τ before confessing is

$$\mathcal{F}(0; \tau, t_0) = \mathcal{A}(\tau) + \frac{1}{r+s+c}[\pi - \alpha v - \phi c T]. \tag{23}$$

Where we use (22) for the constant.

We can now establish the optimal behaviour for a decision maker who receives the private bad-news signal that the product is harmful after the limited liability point: $t_0 \geq T$. The value function's dependence on the waiting strategy τ is explored by:

$$\frac{\partial}{\partial \tau}\mathcal{F}(0; \tau, t_0) = \frac{\partial}{\partial \tau}\mathcal{A}(\tau) = e^{-(r+s+c)\tau}[\pi - \alpha v + (r+s)\phi T]$$

Therefore if $\pi - \alpha v + (r+s)\phi T > 0$ then $\mathcal{F}(0; \tau, t_0)$ is increasing in τ and so the decision maker will never confess after public bad news at $t_0 \geq T$. Whereas if $\pi - \alpha v + (r+s)\phi T < 0$ then $\mathcal{F}(0; \tau, t_0)$ is decreasing in τ and so the decision maker will confess immediately after public bad news at $t_0 \geq T$. This proves the proposition for the case in which the bad news arrives at or after the limited liability point T.

We turn now to analyse behaviour before the limited liability point T.

We derive the terminal value at time T, denoted $F(T)$, if the decision maker has received the private bad news signal before this point and waits until T. Using the optimal behaviour derived above to take the appropriate limit of τ in the continuation value (23) we have

$$F(T) = \begin{cases} \frac{1}{r+s+c}[\pi - \alpha v - \phi c T] & \text{if } \pi - \alpha v + (r+s)\phi T > 0 \\ -\phi T & \text{if } \pi - \alpha v + (r+s)\phi T < 0 \end{cases} \tag{24}$$

We now derive the value function of the decision maker who receives the private bad news signal at $t_0 < T$ and decides to wait until calendar time T before following the optimal strategy reflected in (24). The value function in the continuation region in this case is given in (10). The constant can be derived by the end point condition:

$F(T-t_0;t_0) = F(T)$. This allows us to determine the constant as

$$\mathcal{A}(T-t_0) = \begin{cases} \frac{1}{r+s+c} e^{-(r+s+c)(T-t_0)} \frac{c\phi}{r+s+c} & \text{if } \pi - \alpha v + (r+s)\phi T > 0 \\ -\frac{1}{r+s+c} e^{-(r+s+c)(T-t_0)} & \\ \quad \cdot \left[\pi - \alpha v + (r+s)\phi T - \frac{c\phi}{r+s+c}\right] & \text{if } \pi - \alpha v + (r+s)\phi T < 0 \end{cases} \tag{25}$$

So using (10) the value to the decision maker if she receives the private bad news signal at t_0 and intends to wait until T before proceeding optimally at that point is

$$F(0;t_0) = \mathcal{A}(T-t_0) + \frac{1}{r+s+c}\left[\pi - \alpha v - c\phi t_0 - \frac{c\phi}{r+s+c}\right], \tag{26}$$

with the constant $\mathcal{A}(T-t_0)$ given in (25).

We now derive the value of the decision maker if he receives the private bad news signal at $t_0 < T$ and decides to wait for calendar time $\tau < T - t_0$ to pass before confessing. This analysis has been conducted already with the value to the decision maker at the moment the private bad news signal arrives being given by $\mathcal{F}(0;\tau,t_0)$ given in (14). This value function is quasi-convex in the waiting time τ as established in the discussion around (19). It follows that the optimal strategy of the decision maker is either to confess immediately or to wait until $\tau \nearrow T - t_0$. Waiting for a middle point is dominated.

Confession immediately generates a payoff of $-\phi t_0$. It remains to establish whether it is best to wait until nearly calendar time T or exactly calendar time T. If the decision maker waits to nearly calendar time T then the value when the private bad news signal arrives is, using (14)

$$\begin{aligned} \lim_{\tau \nearrow T-t_0} \mathcal{F}(0;\tau,t_0) = &-\frac{1}{r+s+c} e^{-(r+s+c)(T-t_0)}\left[\pi - \alpha v + (r+s)\phi T - \frac{c\phi}{r+s+c}\right] \\ &+ \frac{1}{r+s+c}\left[\pi - \alpha v - c\phi t_0 - \frac{c\phi}{r+s+c}\right] \end{aligned} \tag{27}$$

If the decision maker waits until T and follows the optimal strategy at that point, then the value was derived in (26). Comparing (26) to (27) we observe that if $\pi - \alpha v + (r+s)\phi T > 0$ then there is a discontinuity in the value function between waiting to nearly calendar time T and waiting to exctly calendar time T. It is more profitable to wait to exctly calendar time T:

$$\pi - \alpha v + (r+s)\phi T > 0 \Rightarrow \underset{\text{From (27)}}{\lim_{\tau \nearrow T-t_0} \mathcal{F}(0;\tau,t_0)} < \underset{\text{From (26)}}{\mathcal{F}(0;T-t_0,t_0)}$$

Whereas if $\pi - \alpha v + (r+s)\phi T < 0$ then there is no discontinuity in the value function between waiting to nearly calendar time T and waiting to exctly calendar time T.

So in all cases the optimal strategy if the private bad news signal is received at $t_0 < T$ is either to confess immediately, or to wait until calendar time T and then proceed optimally as established above. The value from the waiting to T strategy is given by $F(0; t_0)$ as given in (26) with the constant derived in (25). The overall optimal strategy is given by the higher of the waiting value function (26) as compared to the value from immediate confession of $-\phi t_0$.

We show that if $\pi - \alpha v + (r+s)\phi T > 0$ then waiting to T (and therefore waiting indefinitely) is optimal for the decision maker. First note that in this case $\lim_{t_0 \to T} F(0,; t_0) > 0 > -\phi T$. Hence if the bad news comes at a calendar time approaching T then the decision maker will optimally wait until T and not confess immediately. Now we compare the slope of the two value functions and show a single crossing condition:

$$\begin{aligned}
\frac{\partial}{\partial t_0} F(0; t_0) =& \frac{1}{r+s+c}\left[-c\phi + c\phi e^{-(r+s+c)(T-t_0)}\right] \\
&> -\frac{c\phi}{r+s+c} > -\phi \\
&= \frac{\partial}{\partial t_0}(-\phi t_0)
\end{aligned}$$

Therefore there exsts a unique $t_{0*} < T$ such that the decision maker will not confess if the private bad news arrives after t_{0*}.

We show that if $\pi - \alpha v + (r+s)\phi T < 0$ then immediate confession is optimal for the decision maker. If the bad news arrives just before calendar time T then the payoff from immediate confession and from waiting to T and then confession approach each other and equal $-\phi T$. We now compare the slope of the two value functions and show they do not cross:

$$\begin{aligned}
\frac{\partial}{\partial t_0} F(0; t_0) =& \frac{1}{r+s+c}\left[(r+s+c)e^{-(r+s+c)(T-t_0)}\left(-(r+s)\phi T - (\pi - \alpha v) + \frac{c\phi}{r+s+c}\right) - c\phi\right] \\
&> -\frac{c\phi}{r+s+c} > -\phi \\
&= \frac{\partial}{\partial t_0}(-\phi t_0)
\end{aligned}$$

Therefore the value function from immediate confession is higher for all $t_0 \in [0, T]$, yielding immediate confession as best.

This therefore completes the proof. ■

Proof of Proposition 7. First note that V^{good} is independent of the decision maker's ethics as no harm is created if the activity is good.[26] But with probability q the decision

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

maker anticipates that the activity will be harmful, and therefore lock-in is possible if the signal to this effect comes late enough. Proposition 7 therefore follows if we can show that

$$\partial V^{bad}/\partial\alpha < 0. \tag{28}$$

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 of motion is:

$$V^{bad}(t) = (\pi - \alpha v)dt + \frac{1 - sdt}{1 + rdt} V^{bad}(t + dt) \tag{29}$$

Equation (29) yields differential equation,

$$(r + s)V^{bad} - \dot{V}^{bad} + \pi = \alpha v,$$

and this has general solution

$$V^{bad}(t) = \frac{\pi - \alpha v}{r + s} + \mathcal{A}e^{(r+s)t} \tag{30}$$

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 $\lim_{\tau\to\infty}$ (14), 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 (30) yields the constant $\mathcal{A}$. Therefore we can estab-

this setting then yields that

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

lish the value the agent anticipates at the beginning of the game:

$$V^{bad}(0)|_{t_0>t_{0*}} = \frac{\pi - \alpha v}{r+s}\left[1 - \frac{c}{r+s+c}e^{-(r+s)t_0}\right] - \frac{c\phi}{r+s+c}e^{-(r+s)t_0}\left[t_0 + \frac{1}{r+s+c}\right]. \tag{31}$$

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 (30) 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

$$V^{bad}(0)|_{t_0<t_{0*}} = \frac{\pi - \alpha v}{r+s}\left[1 - e^{-(r+s)t_0}\right] - \phi t_0 e^{-(r+s)t_0} \tag{32}$$

Turn now to the arrival times of the private bad news signal. If the practice is harmful then the private bad-news signal arrives as a Poisson process with arrival rate b. The arrival time of the bad news signal in this case is given by the exponential distribution: $f(t) = be^{-bt}$ (Cooper, 2005). The value of the agent at the beginning of the game 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 (31) and (32) 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:

$$\begin{aligned} \frac{\partial V^{bad}}{\partial \alpha} =& -\left(\frac{v}{r+s+b}\right)\left[\frac{1}{r+s} + \frac{b}{r+s+c}e^{-(r+s+b)t_{0*}}\right] \\ &- \frac{\partial t_{0*}}{\partial \alpha}be^{-(r+s+b)t_{0*}}\underbrace{\left(\frac{\pi - \alpha v}{r+s+c} + \frac{\phi}{r+s+c}\left((r+s)t_{0*} - \frac{c}{r+s+c}\right)\right)}_{\mathcal{D}} \end{aligned}$$

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 (28) now follows proving the result. ■

Proof of Proposition 8. Label the distribution of subjective probabilities of harm which can be held by decision makers concerning possible projects as $J(q)$. Write $\mathcal{E}_q(\alpha)$ for the expected probability of harm in the population of launched projects run by decision makers with ethics α: The result follows if $\frac{d}{d\alpha}\mathcal{E}_q(\alpha) < 0$.

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

$$\frac{\partial V^{bad}(\alpha)}{\partial \alpha} < 0 \quad \text{(condition (28))} \quad \Rightarrow \frac{\partial \bar{q}(\alpha)}{\partial \alpha} < 0.$$

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. ■

Proof of Proposition 9. We re-establish the value functions for this new fine regime. Suppose the private bad-news signal arrives at t_0 and the decision maker considers stopping or continuing after a further time t has elapsed. The Bellman equation at this point, so at calendar time $t_0 + t$ since the start of the activity is:

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

The first term of the maximisation is the payoff from voluntarily stopping, the second term is the payoff from continuing. Note that if the public signal arrives then the extra fine ϕ^{extra} is payable as well as the standard fine ϕ.

Taking the limit as $dt \to 0$ the ode which applies in the continuation region is

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

This ode can be solved explicitly by finding a particular and a general solution in the manner of the main model. The solution is:

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

for some constant $\mathcal{A}$.

Now define $\mathcal{F}(t;\tau,t_0)$ the value after time t has elapsed since the private bad news signal at t_0 when the decision maker's strategy is to confess after time τ has elapsed. For $t < \tau$ the decision maker is in a continuation region. The analysis above therefore applies. It follows that:

$$\begin{aligned}\mathcal{F}(t;\tau,t_0) =& \mathcal{A}_\tau e^{(r+s+c)t} + \frac{1}{r+s+c}\left[\begin{array}{l}\pi - \alpha v - c(\phi+\phi^{extra})(t+t_0) \\ -\frac{c(\phi+\phi^{extra})}{r+s+c}\end{array}\right] \text{ for } t \leq \tau \\ \mathcal{F}(\tau;\tau,t_0) =& -\phi(\tau+t_0)\end{aligned}$$

The second line is the moving boundary condition. Value matching at the boundary allows us to determine the constant $\mathcal{A}_\tau$ as

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

The moment that the private bad news signal arrives, a strategy of delaying for a further τ time before confessing yields a value to the decision maker of:

$$\begin{aligned}\mathcal{F}(0;\tau,t_0) =& \frac{1}{r+s+c}\left[\pi - \alpha v - c(\phi+\phi^{extra})t_0 - \frac{c(\phi+\phi^{extra})}{r+s+c}\right] \\ & - \frac{e^{-(r+s+c)\tau}}{r+s+c}\left[\pi - \alpha v + [(r+s)\phi - c\phi^{extra}](\tau+t_0) - \frac{c(\phi+\phi^{extra})}{r+s+c}\right]\end{aligned} \tag{33}$$

The solution to the Bellman equation at the moment that the private bad news signal arrives is therefore:

$$F(0;t_0) = \sup_{\tau \geq 0} \mathcal{F}(0;\tau,t_0) \tag{34}$$

To solve for the supremum in τ we first differentiate $\mathcal{F}(0;\tau,t_0)$, given in (33), with respect to τ. This yields:

$$\frac{\partial\mathcal{F}(0;\tau,t_0)}{\partial\tau} = e^{-(r+s+c)\tau}\left[\begin{array}{l}\pi - \alpha v - \frac{c(\phi+\phi^{extra})}{r+s+c} \\ +[(r+s)\phi - c\phi^{extra}]\left(\tau + t_0 - \frac{1}{r+s+c}\right)\end{array}\right] \tag{35}$$

We now show that if

$$\phi^{extra} < \left(\frac{r+s}{c}\right)\phi, \tag{36}$$

$\mathcal{F}(0;\tau,t_0)$ is quasi-convex in τ. By inspection of (35) there is exactly one stationary point which we label $\tilde{\tau}$ and by construction $\frac{\partial \mathcal{F}(0;\tilde{\tau},t_0)}{\partial \tau} = 0$. Now note for any $\tau < \tilde{\tau}$ we have $\frac{\partial \mathcal{F}(0;\tau,t_0)}{\partial \tau} < \frac{\partial \mathcal{F}(0;\tilde{\tau},t_0)}{\partial \tau} = 0$ using (6). Similarly for any $\tau > \tilde{\tau}$ we have $\frac{\partial \mathcal{F}(0;\tau,t_0)}{\partial \tau} > \frac{\partial \mathcal{F}(0;\tilde{\tau},t_0)}{\partial \tau} = 0$ using (6). Hence quasi-convexity is established.

By analogous reasoning we establish that if condition (7) holds then $\mathcal{F}(0;\tau,t_0)$ is quasi-concave in τ.

We now prove the first part of Proposition 9. Under condition (36), $\mathcal{F}(0;\tau,t_0)$ is established as quasiconvex in τ. Therefore the supremum in τ is attained at one of $\{0,\infty\}$. The decision maker will confess to the activity immediately if a bad news signal arrives at t_0 if and only if

$$\mathcal{F}(0;0,t_0) > \lim_{\tau\to\infty} \mathcal{F}(0;\tau,t_0)$$

And this simplifies down to

$$t_0 < \left[\alpha v - \pi + \frac{c(\phi + \phi^{extra})}{r+s+c}\right] \cdot \frac{1}{(r+s)\phi - c\phi^{extra}} := \tilde{t}_{0*}$$

Where we have labeled the right hand side $\tilde{t}_{0*}$. If we define

$$t_{0*} := \max(\tilde{t}_{0*}, 0) = \max\left(0, \left[\alpha v - \pi + \frac{c(\phi + \phi^{extra})}{r+s+c}\right] \cdot \frac{1}{(r+s)\phi - c\phi^{extra}}\right) \tag{37}$$

then the result follows.

We now prove the second part of the theorem. Under condition (7), $\mathcal{F}(0;\tau,t_0)$ is established as quasiconcave in τ. Denote the global maximum as occurring at $\tau = \tilde{\tau}$ so that $\frac{\partial \mathcal{F}(0;\tilde{\tau},t_0)}{\partial \tau} = 0$. Using (35) we have

$$\tilde{\tau} + t_0 = \frac{1}{r+s+c} - \left[\alpha v - \pi + \frac{c(\phi + \phi^{extra})}{r+s+c}\right] \cdot \frac{1}{c\phi^{extra} - (r+s)\phi} \tag{38}$$

Note that the right hand side of (38) is independent of t_0. Therefore set

$$\begin{aligned} \mathcal{T}_{0*} :=& \max(\tilde{\tau} + t_0, 0) \\ =& \max\left(0, \frac{1}{r+s+c} - \left[\alpha v - \pi + \frac{c(\phi + \phi^{extra})}{r+s+c}\right] \cdot \frac{1}{c\phi^{extra} - (r+s)\phi}\right) \end{aligned} \tag{39}$$

which is identical to the right hand side of (38) when it is positive and so can be interpreted as a calendar time. It follows that the solution to (34) is given by τ such that $\tau + t_0 = \mathcal{T}_{0*}$ if this is possible, or to confess immediately if not. The final part of the result is to establish that

$$\mathcal{T}_{0*} \leq 0 \Leftrightarrow \frac{c\phi^{extra} - (r+s)\phi}{r+s+c} \leq \alpha v - \pi + \frac{c(\phi + \phi^{extra})}{r+s+c} \text{ using (7)}$$
$$\Leftrightarrow 0 \leq \alpha v - \pi + \phi,$$

completing the proof. ■

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, D. J. and J. H. Kagel (2016). Other-regarding preferences. *The handbook of experimental economics 2*, 217.

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.

Newlands, C. (2025). 30 years on: Nick leeson and the fall of the house of barings. *The Banker*.

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