CAPACITY AND CULPABILITY

JAMES MANWARING

BALLIOL COLLEGE

D.PHIL THESIS

Trinity 2019

WORDS: 84 300

Abstract

Capacity and Culpability

James Manwaring, Balliol College

DPhil Thesis, Trinity 2019

How should the criminal law account for defendants' incapacities? It's often claimed that some incapacities make individuals less culpable for wrongdoing. But what follows if this is true? Like many others, I claim that the criminal law ought not to convict offenders disproportionately to their culpability. Thus, if incapacities render individuals less culpable, then the criminal law ought to account for that lowered culpability. But rather than accepting this 'culpability principle' under the guise of non-instrumentalist retributivism, I instead derive it from instrumental considerations regarding fair labelling.

Next, I survey how the law accounts for incapacities. I generate and populate a novel threefold taxonomy of incapacity rules, comprising what I call incapacity doctrines, incapacity relativisations, and counterfactual incapacity relativisations. I then defend the use of these incapacity rules from certain critiques, including from those who argue for the abolition of some incapacity rules. Many incapacity rules are formulated not (only) by reference to their target incapacities, but (also) by reference to certain proxy attributes, and in particular to the relevant incapacity's aetiology. I defend the use of such proxies but critique those rules which require proof of aetiological proxies in addition to proof of the underlying incapacity.

The final chapters ask why and how incapacities exculpate. I defend the widely accepted view that incapacities exculpate if they make defendants less able to conform to the requirements of a norm. Finally, I ask what it means for one to be incapable of something. I argue that we cannot rely only on metaphysical considerations when ascribing incapacities for the attribution culpability.

Acknowledgements

First and foremost, thanks to my supervisor, James Edwards, for the close scrutiny of every argument within this thesis. Almost every passage has been improved in anticipation of or response to those comments (usually both).

Thanks to the AHRC for funding, without which I'd now be a solicitor. Thanks too to the Oxford law faculty and Balliol College for previous financial support.

Thanks to John Gardner and Andrew Simester for their comments when examining the MPhil thesis from which this grew, and to Andrew Ashworth and Kate Greasley for comments on chapters 1 and 2 at confirmation of status.

Thanks to the community around the Oxford Jurisprudence Discussion Group, including Mikolaj Barczentewicz, Leo Boonzaier, Achas Burin, Sam Burke, Hasan Dindjer, Matthew Kruger, Ezequiel Monti, Robert Mullins, and Leah Trueblood for various helpful discussions.

Thanks to Susan Shi for everything.

John Gardner's writing on the philosophy of criminal law hooked me on the subject. It was in part his influence that drew me to Oxford for the BCL (taking all of his classes), and to stay on for research degrees. It was a pleasure to have him as a tutor and then an examiner. In this thesis I cite more of his work, more frequently, than anyone else. He died too young in July 2019. I dedicate this to him.

Contents

Table of cases

Codere (1916) 12 Cr App R 21

Cole [1993] Crim LR 300

Coley [2013] EWCA Crim 223

DPP v Morgan [1976] AC 182

Dowds [2012] EWCA Crim 281

Elliott v C [1983] 1 WLR 939

Emery (1993) 14 Cr App R (S) 394

Fairchild v Glenhaven Funeral Services Ltd [2002] UKHL 22

Foster v CPS [2013] EWHC 3885 (Admin)

Foster's Case (1825) 1 Lew. 187

G [2003] UKHL 50

Gillick v West Norfolk and Wisbech AHA (1985) 3 All ER 402 (HL)

Golds [2016] UKSC 61

Graham [1982] 1 WLR 294

Gray v Jones [1939] 1 All ER 798

Grewal [2010] EWCA 2448

Hardie [1985] 1 WLR 64

Heard [2008] QB 43

Hegarty [1994] Crim LR 353

Hennessy [1989] 1 WLR 287

Hill v Baxter [1958] 1 QB 277

Holden [1991] Crim LR 478

Horne [1994] Crim LR 584

Howe [1987] AC 417

Hughes [2013] UKSC 56

Hurst [1995] 1 Cr App R 82

Ivey v Genting Casinos [2017] UKSC 67

Jameel v Dow Jones [2005] EWCA Civ 75

Jameel v Wall Street Journal Europe Ltd [2006] UKHL 44

Jheeta [2007] EWCA Crim 1699

John v MGN Ltd [1997] QB 586

Jones (1986) 83 Cr App R 375

JTB [2009] UKHL 20

Stephenson [1979] QB 695

Sullivan [1984] AC 156

The Queen against Daniel M'Naughton (1843) 4 St Tr 847

Turner (1974) 60 Cr App R 80

Wade (1869) 11 Cox CC 549

Wilcocks [2016] EWCA Crim 2043

William Press [2013] EWCA Crim 1849

Williams (Gladstone) (1983) 78 Cr App R 276

Woolmington v DPP [1935] AC 462

Youssoupoff v. MGM Pictures Ltd (1934) 50 TLR 581

Z/Hasan [2005] 2 AC 467

Table of statutes

Table of foreign legal sources

Table of official sources

ALI, *Model Penal Code* (1962)

CIA, *World Factbook*

FBI, *Crime in the United States 2016*

— — *Crime in the United States 2017*

HL Deb 5 March 1968, vol 289

— — 7 March 1968, vol 289

— — 8 April 1968, vol 291

Home Office, *Atkin Committee on Insanity and Crime* (Cmnd 2005, 1923)

— — *Report of the Committee on Mentally Abnormal Offenders* [Butler Report] (Cmnd 6244, 1975)

— — Committee Report on Defamation Cmnd 5909 (1975)

Law Commission, *Partial Defences to Murder* (Final Report, 2004)

— — *Murder, Manslaughter and Infanticide* (Law Com No 304, 2006)

— — *Criminal Liability: Insanity and Automatism* (Discussion Paper, 2013)

— — *Unfitness to Plead Vol 2: Draft Legislation* (2016)

Ministry of Justice, 'Criminal Justice Statistics quarterly, England and Wales, April 2016 to March 2017 (provisional)', 17 August 2017

National Policing Improvement Agency Circular, NPIA 02/2011, 'Recruitment eligibility for police constables', available at: <https://web.archive.org/web/20170918154344/http://www.policecouldyou.co.uk/documents/npia-02-20112835.pdf> (archived 18 September 2017)

1 A Culpability Principle

This thesis evaluates rules of criminal law which account for defendants' incapacities. To evaluate the law, we need some standard of evaluation. We could use general moral standards. But the complexity of both the criminal law and morality makes this overwhelming. There are two common ways of dealing with this complexity: to simplify the evaluand, and to simplify the standard of evaluation. I'll simplify the evaluand by focusing primarily on the substantive law. Even within the substantive law, there's little to say that's equally true of murder, speeding, and obscure industry-specific regulatory offences. Thus, I'll follow other theorists in focusing only on the core 'stigmatic' crimes and their defences.[1] That gives us a workable evaluand. I'll simplify the standard of evaluation by specifying an evaluative principle tailored specifically to criminal law, rather than relying on general moral standards. Criminal law theorists have relied on harm principles, offense principles, principles of legal moralism, and others.[2] But those well-known principles are ill-suited to evaluate incapacity-oriented doctrines. That's because the standard explanation for these doctrines is that they give incapable defendants a break because those defendants are not culpable. This standard explanation doesn't invoke harm, offense, or wrongfulness. The task for this chapter, then, is to specify a *culpability principle* to evaluate the core stigmatic crimes.

Talk of a culpability principle could confuse for two ways. First, the word 'culpability' is used in different ways. Criminal lawyers sometimes use 'culpability' to refer to the fault element of

[1] This division is commonplace in criminal law theory: Andrew von Hirsch, *Censure and Sanction* (Oxford 1993) 7 (on 'the core conduct with which the criminal law deals'); Andrew Simester, 'Is Strict Liability Always Wrong?' in Andrew Simester (ed), *Appraising Strict Liability* (Oxford 2005), discussed by John Stanton-Ife, 'Strict Liability: Stigma and Regret' (2007) 27 OJLS 151; Andrew Ashworth, 'Ignorance of the Criminal Law, and Duties to Avoid It' (2011) 74 MLR 1, 8-10, 12-13 (using the distinction, but accepting that it cannot easily be mapped to legal doctrine); Andrew Ashworth and Jeremy Horder, *Principles of Criminal Law* (7th edn, Oxford 2013) 1-2, 17 (implicitly acknowledging the distinction).

[2] I defend this approach in James Manwaring, 'Criminal Law Principles' (draft), in response to criticisms from Victor Tadros, 'How Not to Think About Criminalisation I' in *Wrongs and Crimes* (Oxford 2017).

crimes, or as a synonym for being found guilty in a criminal trial.[3] By contrast, I'll use 'culpability' to mean *moral blameworthiness*.[4] I leave open what *grounds* moral blameworthiness.[5] Most theorists accept that culpability qua blameworthiness is important when evaluating the criminal law.[6] But the second potential confusion is that this is rarely cashed out in terms of a 'culpability principle'. Other labels are used. For instance, legal moralists (sometimes) hold that culpable wrongdoing is relevant to permissible criminalisation.[7] But legal moralism's focus is on the wrongdoing rather than the culpability.[8] As culpability is distinct from wrongdoing, we shouldn't treat them together.[9] Similarly, *retributivists* hold that a defendant's culpability affects the justifiability of criminal punishment. But, while many theses fly under the banner of retributivism, a core commitment is often said to be its *desert claim*: the idea that punishment (or suffering) can be deserved.[10] The desert claim is highly controversial; much more so than many kinds of culpability principle.[11] It's generally inadvisable to rely on more-controversial

[3] Eg Ashworth refers to any mens rea element as a 'culpability requirement': Ashworth, 'Ignorance of the Criminal Law' (n 1) 7.

[4] While some refer to a notion of specifically *legal* or *criminal* culpability, a major component of that is always moral blameworthiness. My argument goes through *mutatis mutandis*. Mark Dsouza, 'Criminal Culpability after the Act' (2015) 26 Kings LJ 440; Alexander Sarch, 'Who Cares What You Think? Criminal Culpability and the Irrelevance of Unmanifested Mental States' (2017) 36 Law and Philosophy 707, 709; cf Gideon Yaffe, *The Age of Culpability* (Oxford 2018) ch 3.

[5] The most popular view is that blameworthiness involves insufficient regard for others.

[6] 'Many criminal theorists agree that conviction of a…stigmatic criminal offence…should not normally occur unless the accused is culpable with respect to that offence': Andrew Simester, 'A Disintegrated Theory of Culpability' in Denis Baker and Jeremy Horder (eds), *The Sanctity of Life and the Criminal Law: The Legacy of Glanville Williams* (Cambridge 2013) 178.

[7] Legal moralists 'share a thesis about a connection between *culpable* wrongdoing and deserved punishment': Douglas Husak, 'What's Legal about Legal Moralism?' (2017) 54 San Diego L Rev 381, 382 (emphasis added).

[8] According to Anthony Duff's 'modest' legal moralism, we are answerable even for non-culpable public wrongs. (Of course, our answer may be exculpatory). RA Duff, 'Towards a Modest Legal Moralism' (2014) 8 Crim Law and Philos 217.

[9] This point is stronger yet if culpability can exist absent wrongdoing, eg by doing the right thing for culpably wrong reasons. (Eg perhaps a lifeguard acts culpably for using race to select which of two drowning children to save, even if she had sufficient reason to save that child and thus (arguably) committed no wrong in doing so). But nothing turns on this.

[10] Douglas Husak, 'Retributivism in Extremis' (2013) 32 Law and Philosophy 3; Douglas Husak, 'Broad Culpability and the Retributivist Dream' (2012) 9 *Ohio St J Crim* 449. Mitchell Berman points out the necessity of relying on *suffering* as the thing-to-be-deserved in 'Two Kinds of Retributivism' in RA Duff and Stuart Green (eds), *Philosophical Foundations of Criminal Law* (Oxford 2011) 437-438.

[11] Hart famously referred to the desert claim as involving a kind of 'moral alchemy' in *Punishment and Responsibility* (2nd edn, Gardner intro, Oxford 2008) 234-5, echoed by Victor Tadros, *The Ends of Harm* (Oxford 2011) ch 4, esp. 73-54. Tadros objects that it's not clear why suffering (see *ibid*) is deserved as opposed to some other negative outcomes that few retributivists would accept are legitimate modes of punishment, such as destroying valuable relationships or self-knowledge.

premises for less-controversial conclusions. My task, modified, is to specify a plausible culpability principle without relying on the desert claim.[12] My claim is that we can derive a desert-free culpability principle from the idea of *fair labelling*.

1 Fair labelling

Andrew Ashworth coined the idea of fair labelling in these terms:

> *FL*: 'the label applied to an offence ought fairly to represent the offender's wrongdoing'[13]

A principle of fair labelling has since been widely supported by criminal law theorists.[14] But while Ashworth (and others) formulated the principle in terms of labelling *wrongdoing*, this understates the intuitive force of the claim. For we should also fairly represent the *culpability* of that wrongdoing. There are two routes to this extension. First, by 'wrongdoing' we might mean a breach of a duty, defined as an act-type (such as a duty not to kill).[15] Then *FL* tells us to distinguish this act-type from others (like theft). But *within*-act-type differences can be more morally significant than *between*-act-type differences: compare negligent with intentional killings. The significance of many within-act-type differences is explained in terms of culpability. Thus, fairly labelling culpability is at least as important as fairly labelling wrongdoing.[16] Second, by 'wrongdoing' we might already mean to incorporate culpability. On this view, accidental and intentional killings are different types of 'wrongdoing' because they

---

[12] Retributive culpability principles tend to focus on punishment as the relevant evaluand within the substantive criminal law. By contrast, I will focus on conviction as the relevant stage, for reasons developed below, and by Simester, 'Is Strict Liability Always Wrong?' (n 1) 27.

[13] Andrew Ashworth, 'The Elasticity of Mens Rea' in CFH Tapper (ed), *Crime, Proof and Punishment: Essays in the Memory of Sir Rupert Cross* (Butterworth 1981) 53. Ashworth actually coined this 'representative labelling', but 'fair labelling' stuck.

[14] James Chalmers and Fiona Leverick offer a historical overview of the principle and cite the assent of various theorists, textbooks, and law reform committees: 'Fair Labelling in Criminal Law' (2008) 71 MLR 217. Unfortunately, as Chalmers and Leverick note, the broad support among theorists for a principle of fair labelling is not matched by an equivalent depth of analysis. Aside from their article, I know only of one devoted to the principle: Victor Tadros's 'Fair Labelling and Social Solidarity' in Lucia Zedner and Julian Roberts (eds), *Principles and Values in Criminal Law and Criminal Justice Essays in Honour of Andrew Ashworth* (Oxford 2012). (Though at the time of writing Matt Gibson has forthcoming work on the principle).

[15] John Gardner, 'In Defence of Defences' in *Offences and Defences* (Oxford 2007), discussed by Andrew Simester, 'Wrongs and Reasons' (2009) 72 MLR 648.

[16] An intentional killer is plausibly morally closer to an intentional fraudster like Bernie Madoff than a negligent killer like a tired doctor who causes the death of a patient.

are culpable to different degrees. For whichever reason, Ashworth himself extends the idea to culpability.[17] Thus *FL* implies a *Fair Labelling Culpability Principle*:

> *FLCP1*: offence labels ought fairly to represent offenders' *culpability*

But this formulation already seems too restrictive. Offence labels don't capture all culpability-relevant features of a defendant's conduct. The offence of battery is committed if D intentionally hits a non-consenting V. But committing the offence doesn't make D criminally liable or culpable without further ado. D may have hit V in justified self-defence. It would be wrong to label D a 'batterer' (implying culpability) without mentioning her defence.[18] Hence:

> *FLCP2*: *crime* labels ought fairly to represent offenders' culpability

'Crime labels' (excuse the inelegant phrase) account for both offences and defences. Because 'offenders' are only 'represented' by crime labels if they are *convicted* of a crime, we can clarify *FLCP2* to read:

> *FLCP*: people ought to be *convicted* of crimes only insofar as the crime label fairly represents their culpability

FLCP is close to the sort of culpability principle we're interested in. Something like *FLCP* also has some (tentative) legal support.[19] But what's the *argument* for *FLCP*?

Andrew Simester suggests one argument. Convictions for stigmatic crimes entails that

---

[17] Ashworth originally pitched fair labelling in contradistinction to 'a subjectivist principle that an individual's moral culpability was fairly reflected in his sentence'. But shortly thereafter he notes the Criminal Law Revision Committee's view that people labelled should be 'deserving of the stigma', and claims that 'the legal designation of an offence should fairly represent the nature of the offender's *criminality*': 'The Elasticity of Mens Rea' (n 13) 53-56. This extension becomes more explicit later. Ashworth claims that D 'does not deserve…a heavily condemnatory label when his culpability was much more minor.' Andrew Ashworth, 'Taking the Consequences' in Stephen Shute, John Gardner, and Jeremy Horder (eds), *Action and Value in Criminal Law* (Oxford 1993) 118. (While Ashworth presents this as a 'subjectivist's' claim, he endorses that position, eg at 124). Most explicitly, he has claimed that 'criminal conviction involves public censure, and that should be limited to cases of culpable wrongdoing', that it is a fundamental problem to subject 'people to the censure of conviction when they may not have been culpable', that 'it is unjustifiable for the State to provide for the conviction of people who are not at fault' and that 'the offences [the State] creates should be so framed as to impose the censure of conviction only for culpable wrongdoing'. Ashworth, 'Ignorance of the Criminal Law' (n 1) 4-5, 12, 20.

[18] Chalmers and Leverick, 'Fair Labelling in Criminal Law' (n 14) 244-246.

[19] '[I]t is a salutary principle that conviction of serious crime should depend on proof not simply that the defendant caused...an injurious result to another but that his state of mind when so acting was culpable': *G* [2003] UKHL 50 [32] (Lord Bingham).

> D is labelled as a particular sort of criminal (a "μer"), a labelling that conveys a public implication of culpable wrongdoing… A conviction for μ-ing has the effect of naming D a criminal (in respect of that particular offence), a branding which is communicated to society as well as to D. Assuming that, if imposed on a strict liability basis, the label 'criminal-μ,' continues to retain its stigmatic quality, this amounts to systematic moral defamation by the state. Given the public understanding of that designation, when it labels him a criminal the state is no longer telling the public the truth about D. People have a right not to be censured falsely as criminals, a right that is violated when one is convicted and punished for a stigmatic crime without proof of culpable wrongdoing.[20]

The use of strict liability is but one way that crime labels may fail to 'tell the truth' about defendants, resulting in false censure.[21] According to Simester, labels are truthful only if they convey a level of culpability commensurate to the public understanding of the act in question. Strict liability crimes are an extreme case, where defendants may be *entirely* non-culpable ('morally innocent').[22] But the same objection applies, mutatis mutandis, to cases where defendants are convicted *disproportionately* to culpability. Strict liability is simply an extreme case of potentially disproportionate-to-culpability conviction. Hence Simester's argument generalises to support something like *FLCP*: people ought to be convicted of crimes only insofar as the crime label fairly represents their culpability.

Simester's argument is capable of many interpretations. The best, I think, is:

Defamation Argument

(1) Conviction for stigmatic crimes labels offenders as culpable wrongdoers
(2) Offenders who commit stigmatic crimes are culpable to some degree X
(3) The public believe that crime labels convey offender culpability to degree Y

[20] Simester, 'Is Strict Liability Always Wrong?' (n 1) 33-34.

[21] Simester treats his defamation argument separately from representative (ie fair) labelling, which, he thinks, is responsive to *harm* (ibid 46). We are only concerned with fair labelling qua culpability, and thus can treat his defamation argument as a subset of (his interpretation of) fair labelling views.

[22] 'Strict liability leads to conviction of persons who are, morally speaking, innocent. Convicting and punishing those who do not deserve it perpetrates a serious wrong', and 'Morally speaking, it is wrong to convict the innocent.' ibid 21, 34.

(4) It is morally defamatory to use crime labels where the public belief in the degree of offender culpability which attaches to that label (Y) exceeds offenders' actual culpability (X), ie to use crime labels where Y>X.[23]
(5) Moral defamation is morally objectionable
(6) Therefore, it is morally objectionable to use crime labels where the public belief in the degree of culpability which attaches to that label exceeds offenders' actual culpability

Premise (1) is reasonably uncontroversial.[24] Premise (2) should be uncontroversial, given that X can take the value 0, ie 'not at all culpable'. (Per Simester's example of the 'morally innocent' defendant for strict liability crimes.) Premise (3) only requires details as to what count as public beliefs.[25] The interesting part of the argument is premises (4) and (5) and the conclusion (6). I'll defend the premises in this and the next section and consider the strength of the conclusion in the final section.

Consider first *why* moral defamation is morally objectionable, premise (5). Simester distinguishes

> between censure, which the state expresses through its *action* of convicting (and punishing) the defendant, and the *effect* of that action, in terms of the stigma that attaches to D and his conduct… [O]ne reason why the state ought not falsely to censure D for a serious crime is supplied by the consequences for D's life. But the two do not always go together, and D has a right to be neither falsely censured nor

[23] Below I use subscripts to identify culpability magnitudes (eg X_{10}). This is illustrative only: I'm not claiming that culpability can be precisely quantified. Magnitudes will be vague at best. This vagueness means that the X/Y differential must be quite severe to be confident of no overlap. The discrepancy must be correspondingly severe to count as defamatory. Again, I don't claim to specify precisely how severe.

[24] Andrew Cornford claims that 'we have no particular reason to suppose that [the label 'manslaughter'] implies culpability as to death. How, indeed, could we even determine whether this were true?... Just because a label attributes a result to an actor, it does not follow that it also implies culpability as to result.': 'The Architecture of Homicide' (2014) 34 OJLS 819, 827. While this is certainly an empirical question, I find it very plausible that crime labels imply culpability as to results. All I require, however, is that crime labels imply '*some* culpability', which Cornford accepts is the case.

[25] It would be implausible to require that *every* member of public share some view. Rather, a 'public' belief must be held by some subset or subsets whose beliefs affect D. Further, their belief should generally be *common*: they believe *that the others believe* that crime label *L* conveys culpability *N*. There are more complicated cases: a relevant subset could instead believe *that there is a relevant subset* with the relevant beliefs. This false common belief provides *apparent* reason to shun the defendant around which they could coordinate to affect D. There are yet more complex possibilities: further possible recursions; discontinuities between the propositions, including negatives, etc. But they will rarely be practically relevant. I discuss the possibility of picking out a *reasonable* subset at §2.5.

> falsely stigmatized. Even if D suffers no stigma, the state should not purport to censure him without believing him to be culpable. Telling lies is wrong in itself and not merely because of the consequences.[26]

Call the objection to the '*action* of convicting'—that it is 'wrong in itself' to falsely censure—the *intrinsic objection*.[27] Call the objection to the 'effect of that action'—that it stigmatises the defendant, resulting in bad consequences—the *instrumental objection*. The intrinsic objection offers a simple and powerful explanation of moral defamation's objectionability. It is simple and powerful because it doesn't rely on the instrumental, contingent consequences of unfair labelling. But it is precisely this non-contingency that makes the intrinsic objection either implausible or unhelpful.

It's implausible to think that lying is always wrong *all-things-considered*, that is, unjustifiable.[28] The criminal law doesn't tell the truth about many things: it uses aliases for certain defendants and victims, it deems away certain types of evidence, and it calls defendants 'not guilty' even if everyone is confident (but not beyond a reasonable doubt) that they are guilty. These untruths are not wrong in themselves. They are, we think, justified half-truths or white lies.

The more plausible interpretation of the intrinsic objection is that lying is always pro tanto wrong, ie a breach of a duty. Even here, however, we may doubt that there is any reason (a fortiori, a duty) not to lie when no bad consequence would result.[29] Is this plausible for *morally defamatory* lies?[30] That depends on what it takes for a lie to be morally defamatory. One might

---

[26] Simester, 'Is Strict Liability Always Wrong?' (n 1) 34-35 (citations omitted).

[27] 'It is…arguable that there are instrumental benefits to be gained from the device of strict liability… However, assuming they exist, those benefits must be weighed against the intrinsic moral objections to strict liability… In the context of stigmatic crimes, it seems to me that these objections are decisive.' ibid 33.

[28] Section 3, below. While Kant famously demurs in 'On a Supposed Right to Lie from Altruistic Motives' in *Practical Philosophy* (Gregor trs, Cambridge 1996), several Kant scholars reject or caveat that interpretation: Robert Benton, 'Political Expediency and Lying: Kant vs Benjamin Constant' (1982) 43 JHisIdeas 135; Christine Korsgaard, 'The Right to Lie: Kant on Dealing with Evil' (1986) 15 Philosophy and Public Affairs 325; Helga Varden, 'Kant and Lying to the Murderer at the Door... One More Time: Kant's Legal Philosophy and Lies to Murderers and Nazis' (2010) 41 JSocPhil 403.

[29] I assume throughout that the liar intends the actual outcome. It could still be wrong to *attempt* to lie with harmful consequences that don't eventuate, or (perhaps) *negligently* to cause harm despite lacking an intention to do so.

[30] Simester claims that moral defamation via conviction is 'no ordinary lie' and, unlike ordinary defamation, it is 'imposed by the state' in an 'authoritative voice' that 'alienate[s] D from society'. But the examples I gave in the text were all lies by the state, using the criminal law, that are nonetheless not intrinsically wrongful. Simester obliquely acknowledges this, as he elaborates the objection in instrumental terms: the conviction 'marks D out in such a way that it becomes appropriate, within the

plausibly claim that lies are defamatory only if they cause harm to reputation. Indeed, this is the position in tort.[31] If lies are wrong because morally defamatory, and morally defamatory only if harmful (because of their 'stigmatic effect'), then the wrong of morally defamatory lies is explained by (contingent) harmfulness rather than (intrinsic) wrongfulness.[32] If a public authority wants to lie about me for my benefit, with no bad consequences down the line, I wouldn't call that even *pro tanto* wrong. The wrong of lying seems to reduce to the wrong of harming.[33] The intrinsic objection reduces to the instrumental objection.

Defenders of the intrinsic objection have at least two possible responses. First, they might point to ways in which lying is wrong besides harmful consequences. In commodification debates, many argue that the social meaning of market exchange means that it would be wrong to sell organs or babies even if those sales resulted in good consequences. Analogously, we might think that the social meaning of conviction is sufficiently serious that lying about it is wrong even if good consequences resulted. But I doubt that the social meaning of either market exchange or conviction gets us to intrinsic wrongfulness. As Jason Brennan and Peter Jaworski point out,

> There is little essential meaning to market exchanges or money. What market exchanges mean depends upon a culture's interpretative practices… [T]hese interpretative practices are themselves subject to moral evaluation… [C]ultures sometimes impute meaning to markets in harmful, socially destructive ways. Rather than giving us reason to avoid those markets, it gives us reason to revise the meaning

community' to treat D differently. 'Certain exclusions, both social and professional, may legitimately follow…undermin[ing] D's participation in…society itself.' Simester, 'Is Strict Liability Always Wrong?' (n 1) 34-35.

[31] A lie is defamatory if it will tend to lower a person in the estimation of right-thinking members of society, causing serious harm to reputation. The harm requirement is from the Defamation Act 2013 s1. While the law presumes reputational harm in written libels (they are 'actionable per se'), it would be an abuse of process to institute proceedings where there is only minimal harm: *Jameel v Dow Jones* [2005] EWCA Civ 75, *Jameel v Wall Street Journal Europe Ltd* [2006] UKHL 44 (in which only five people saw a would-be defamatory claim). Slanders (oral/transient defamatory statements) are actionable per se where they allege criminal conviction, but that is traditionally justified on the basis that such slanders lead to social ostracism–another *effect* of the false censure: *Gray v Jones* [1939] 1 All ER 798.

[32] The law of defamation's harm requirement might not imply (1) Only harmful lies *are wrongful*, but instead (2) Only harmful lies *justify the imposition of legal liability*. While I claim (1) here, (2) suffices for my argument (as discussed below). Indeed, I (tentatively) endorse both.

[33] This view of lying has parallels with Patrick Atiyah's view of promising. (Though Atiyah's view is less plausible, for reasons explained by Joseph Raz, 'Promises in Morality and Law' (1982) 95 Harv LR 916 §2.)

> we assign to these markets or, if we can't, to conscientiously rebel against or ignore the meaning our society attaches to these markets.[34]

The same argument applies, mutatis mutandis, to the social meaning of conviction. Say it turned out that certain crime labels resulted in excellent consequences for the defendant. In that world, we should criticise *arguments that objected to* such labels rather than criticising the labels themselves.

A second response is that even if the right not to be falsely censured (the intrinsic objection) is *explained* by harmful consequences (the instrumental objection), it is not *reducible* to it. Consider swearing. The reason why any word is considered taboo, a swear-word, is usually explained by the word's offensive connotation. That swearing causes offense explains why it is wrong to swear. But though offense-causing explains the wrong of swearing, that wrong might persist even in cases where no offense is taken. The wrong of swearing thus has an independent life from the wrong of causing offense.[35] Similarly, we might think that the wrong of falsely censuring has a separate life from the wrong of stigmatisation (the bad consequences) *even if* its wrongfulness is explained by the harm involved in stigmatisation. It might be intrinsically wrong for the state both to curse at us and to censure us, even if we happen to be thick-skinned enough to take no offense or for whatever reason suffer no stigma. This is a better response, but it faces two problems.

First, false censure's independently wrongful life is minimal. Swearing is frequently not wrong at all: it often just adds emphasis or signals comfortable familiarity with another. The cases where swearing remains wrong tie very closely, if not precisely, to the cases where it causes offense. The same applies to the wrong of false censure: insofar as it has an independent life from the wrong of stigmatisation, it is a minimal life. Moreover, even if it is pro tanto wrong to lie, this doesn't tell us *how* wrong. The pro tanto wrong could be easily justified. In practice, the moral force of our objection to lying will depend on how wrong it is, and that issue will

[34] Jason Brennan and Peter Jaworski, 'Markets without Symbolic Limits' (2015) 125 Ethics 1053, 1057-1058. I wouldn't go so far as Brennan and Jaworski: we might have *some* reason to act even in 'socially destructive ways' (eg if acting otherwise would be counterproductive, eg if conscientious objection would result in even worse social destruction). The point is simply that it would take a lot more to get to an intrinsic wrong.

[35] This picture of wrongs follows the picture of rules in Frederick Schauer, *Playing by the Rules* (Oxford 1991).

largely reduce to the question of harm. That is why even the pro tanto version of the intrinsic objection is unhelpful.

Second, the intrinsic objection risks redundancy. It claims that it is 'wrong in itself' to morally defame people, that it is morally defamatory to convey false censure, and that it is falsely censuring to convict disproportionately to culpability. But couldn't we have just said that it's 'wrong in itself' to convict disproportionality to culpability and drop the explanation in terms of false censure and moral defamation?[36] The mere fact that extra terms are added—defamation, false censure—is insufficient to count as a good explanation.[37] To explain what counts as defamation, the defamation argument should be able to tell us something about the limits or extent of these claims.[38] The instrumental objection provides precisely such details: defamation is wrong insofar as it harms the defendant. But the intrinsic version seems extensionally identical to the conclusion that convicting disproportionately to culpability is wrong. These problems suggest that we should focus on the instrumental objection.

The instrumental objection is straightforward: moral defamation via unfair labels stigmatises the defendant, which results in bad consequences. That is why it is morally objectionable. Consider a simple case:

> *Job Application*: You apply to join the police. You're an excellent candidate. Due to a technical glitch, the criminal record check labels you a murderer. As a result, you do not get the job.

Your situation in *Job Application* is patently unfair. The inaccurate and misleading crime label results in unfair bad consequences: you don't get the job; the police lose a good candidate. Fair labelling cases are analogous. Consider:

---

[36] This is not to say that these further claims about false censure and moral defamation themselves need to be explained by deeper claims. Every argument needs premises. The point is just that a purported explanation needs to add something.

[37] Compare the empty 'explanation' of fire in terms of phlogiston.

[38] One objection to negative retributivism is that it purports to explain our objection to punishment in terms of an extensionally identical concept of negative desert. Contrast this empty explanation with Victor Tadros' duty view of punishment. Like the negative retributivist, Tadros accepts that it is wrong to punish disproportionate to culpability. But he explains this (via a mere means principle) in terms of more fundamental claims about duties of rescue and liability to defensive harm, which in turn explains certain limits of punishment for purposes of general deterrence. Tadros, *The Ends of Harm* (n 11) Part IV.

> *Job Application 2*: You apply to join the police. You're an excellent candidate. Your criminal record check *accurately* labels you an assaulter. The police believe you viciously beat your victim. As a result, you do not get the job.[39] In fact, you only gently pushed your victim.

The label 'assaulter' applied to you is not inaccurate. But it is misleadingly overbroad: it covers everything from minor pushing through to vicious beatings.[40] This misleading overbreadth again leads to bad consequences: a good candidate missing out on a job. These bad consequences are not limited to employment opportunities. The stigma associated with both the label 'assaulter' and the generic label 'criminal' will likely bring in tow all sorts of negative societal responses. If the public believe, on the basis of the label, that the labelled offender is culpable to some degree, Y, then they will treat the offender in accordance with such a Y-er. This is plausibly defamatory where the offender was only culpable to some lesser degree X. That is, Y>X. The bad consequences which follow from that mismatch both explain why the mismatch is defamatory, and why it is morally objectionable. This concern underlies objections to unfair labelling in terrorism, homicide, and sex offences.[41] In each case the underlying crime can involve less-culpable conduct (or less proof of it) than the highly stigmatic labels would imply. This unfair labelling is defamatory.

2 Five objections

There are (at least) five interrelated objections to the instrumental moral defamation argument. First, labels can be misleading for reasons other than failing to track culpability. So how do culpability-tracking labels solve the problem? Second, instead of culpability-tracking labels, why not use bland labels? Third, isn't our true objection to unfair convictions, not labellings?

---

[39] The police would be permitted but not required not to hire you. Murder results in automatic disqualification. Assault occasioning ABH or maliciously inflicting GBH result in disqualification subject to exceptional circumstances. See the National Policing Improvement Agency Circular, NPIA 02/2011, 'Recruitment eligibility for police constables', available at: <https://web.archive.org/web/20170918154344/http://www.policecouldyou.co.uk/documents/npia-02-20112835.pdf> (archived 18 September 2017).

[40] Hence why English law contains a variety of offences against the person.

[41] Jacqueline Hodgson and Victor Tadros, 'How to Make a Terrorist Out of Nothing' (2009) 72 MLR 984 (arguing that the Terrorism Act 2000 s58(1) permits the unfair labelling of 'terrorists' based on innocent, everyday conduct); James Edwards, 'Justice Denied: The Criminal Law and the Ouster of the Courts' (2010) 30 OJLS 725, 730 (arguing that the Terrorism Act 2006 s1(2) permits the unfair labelling of 'terrorists' where no such thing had been proved in court). Chalmers and Leverick, 'Fair Labelling in Criminal Law' (n 14) provide further examples.

Fourth, what if bad consequences probably wouldn't arise? Fifth, bad consequences result from the public reaction, not the law. What if the public are unreasonable?

2.1 Other unfairnesses

Bad consequences may arise from misleading labels even if those labels are not misleading as to culpability. Consider:

> *Job Application 3*: You apply to join a law firm. You're an excellent candidate. Your criminal record check accurately labels your conviction of a 'property offence'. The firm believe you engaged in dishonest theft. As a result, you do not get the job. In fact, you recklessly failed to protect client money.

This crime label is misleadingly overbroad. It leads to bad consequences: the firm's objection is to dishonesty offences, not to reckless accounting.[42] And this plausibly grounds a complaint of unfair labelling. But let us stipulate that the label was not misleading as to your culpability. Your reckless accounting was equally culpable as the dishonest theft that the firm envisioned. Thus, a *culpability-tracking* fair labelling principle would not impugn your label nor prevent those bad consequences. Does this not imply that our objection to unfair labelling is not fully explained by the instrumental objection? It does indeed. But this is no objection. Culpability-mismatch cases explain why being accurate *about culpability* matters. But these cases are only a subset of possible unfair labellings. Different mislabellings could lead to different bad consequences and require different solutions.[43] I don't claim that culpability-tracking labels assure fairness across the board.

2.2 Bland labels

We shouldn't worry that culpability-tracking labels don't solve all cases of unfair labelling. But we should worry if culpability-tracking labels are not a good solution *even in cases of moral defamation* (Y>X mismatches). Descriptive labels like 'murder', 'assault' etc will likely retain hard-to-resist Y>X discrepancies no matter how assiduously they try to track culpability. But there is an alternative. The law could use *bland* labels like 'Crime 21'. The public won't have strong views about bland labels. If anything, they'd have to work out exactly how culpable each offender happened to be. Thus Y (publicly perceived offender culpability) would rarely

[42] Dishonesty offences are grounds for being struck off the roll by the Solicitors Regulation Authority.

[43] Ashworth's *wrongdoing*-tracking fair labelling principle could solve *Job Application 3*.

exceed X (actual offender culpability). It seems like the defamation argument supports bland labels rather than culpability-tracking labels.[44]

But consider two doubts about bland labels. First, some worry that bland labels lack the 'moral resonance' of traditional descriptive labels.[45] We could lament this for its own sake. But we can also cash this out in instrumental terms. Descriptive crime labels use terminology that citizens already understand.[46] They provide intuitive guidance as to when criminal liability will arise. The rule-of-law benefits of this moral clarity (as John Gardner calls it) almost certainly exceed any amount of textual tinkering.[47] Bland labels put this guidance at risk.[48]

Second, I doubt that bland labels are a viable proposition. This is for two reasons. One: the public catch up to novel terms. Consider the public understanding of the terms 'section 28' regarding gay rights,[49] being 'sectioned' for mental instability,[50] or the widespread use of disability-related terms as insults.[51] Bland euphemisms quickly come to have all the vivid implications of the euphemised term. Denotation yields to connotation.[52] 'Crime 21' would quickly attach the opprobrium of whatever it replaced. Two: bland labels probably *couldn't* replace descriptive labels. We need to know, at least in broad terms, what sort of conduct we are criminalising, charging, and convicting. Bland labels *instead of* descriptive labels would result in *unclear* labels. Worse, for want of descriptive labels to discern between less and more culpable

---

[44] Recall *FLCP:* people ought to be convicted of crimes only insofar as the crime label fairly represents their culpability

[45] Cornford, 'The Architecture of Homicide' (n 24) 835-838, citing the Law Commission, *Murder, Manslaughter and Infanticide* (Law Com No 304, 2006) 22, which took the maintenance of this moral resonance to be the primary reason to retain the traditional offence labels in homicide.

[46] John Gardner, 'Rationality and the Rule of Law in Offences Against the Person' in *Offences and Defences* (n 15) 45.

[47] Cf Ashworth, 'Ignorance of the Criminal Law' (n 1).

[48] While plausible, it's ultimately an empirical matter whether descriptive labels are clearer than alternatives. I'm less confident than Gardner that 'GBH' is clearer than 'serious harm'. I don't know whether Canada's recent abolition of the specific crime of 'rape' has had positive or negative consequences. For this reason, I place more weight on the second objection, below.

[49] The Local Government Act 1988 s28 prevented local authorities from, inter alia, 'promot[ing] homosexuality' (since repealed by the Local Government Act 2003 s122).

[50] Ie involuntarily detained per section 2 or 3 of the Mental Health Act 1983.

[51] For intellectual disability this includes 'idiot', 'cretin', 'moron', 'imbecile', 'feeble-minded', 'retard', 'special' and 'autistic'. For physical disabilities they include 'cripple', 'invalid', 'spastic', etc.

[52] 'Euphemism' in WVO Quine, *Quiddities* (Harvard 1987) 53-54. Steven Pinker dubbed the phenomenon the 'euphemism treadmill' in 'The Game of the Name' *New York Times* (New York, 5 April 1994).

wrongdoers, the public might look with enhanced suspicion at anyone convicted of anything, worsening Y>X discrepancies. But if bland labels are used *together with* traditional descriptive labels it's hard to imagine their survival as anything other than vestigial bureaucratise: the descriptive labels would retain primacy. If we can't avoid descriptive labels playing an important role in the public perception of crimes, culpability-tracking labels are the only obvious solution to Y>X discrepancies.

2.3 The true objection?

Does this miss the point? The defamation argument objects to mislabelling offenders. If correct, this implies that we should not convict disproportionately to culpability. But that implication—that we shouldn't convict disproportionately to culpability—seems intuitively more fundamental than our objection to mislabelling. Wouldn't it remain wrong to convict disproportionately to culpability *even if* mislabelling wasn't wrong? If so, the defamation argument doesn't capture our true objection to conviction disproportionate to culpability.

Happily, I need not claim that the instrumental defamation argument explains every facet of our intuitive objection to disproportionate conviction. There will surely be other reasons in play. Perhaps the most intuitive objection relies on the desert claim. But, as I said at the outset, the desert claim is also highly controversial. My task is to find an alternative argument for a culpability principle. The instrumental defamation argument provides just such an alternative. It may be less intuitive, but it also has countervailing theoretical virtues, including improved explanatory power as to the limits of a culpability principle. That, I think, suffices to justify pursuit of the alternative argument.

2.4 Contingency

The instrumental defamation argument objects to labels that (1) result in Y>X mismatches, which (2) result in bad consequences. Those consequences are contingent. If no bad consequences eventuated our culpability principle would be *inert*. While figuring out the consequences is ultimately an empirical matter, the defamation argument is interesting because bad consequences are highly plausible. The contingency objection is that bad consequences are implausible.

A weak version of this objection can point to the contingency of crime labels' denotation. I claimed that bland labels wouldn't work because the public would quickly learn that 'Crime 21' meant 'murderer', such that bland labels would quickly come to have all the connotations

of descriptive labels. But this euphemism treadmill would also apply to mismatched labels. The label 'murderer' results in bad consequences because it denotes culpable killers. Applying that label to speeders would be unfair because those speeders might then suffer from public reactions more appropriately levelled at culpable killers. That violates our culpability principle. But the public would soon learn to distinguish killer-murderers from speeding-murderers. They would recognise that 'murderer' had two very different senses. It would become a strange homonym with a mild and a serious sense. The public would stop treating speeding-murderers like killer-murders. Thus: even egregiously misleading labels would not result in bad consequences. That's the objection. But adapting to misleading descriptive labels is altogether harder than adapting to euphemistic bland labels. Many law students struggle to distinguish subtle differences between 'assault', 'battery', and 'assault occasioning ABH'. Laypeople have no means of separating less from more culpable tokens of various crime labels. 'Distributor of child pornography' is an aptly stigmatic crime label, but not when applied to children who share explicit images they have taken of themselves. Presented with that label, how could the public discern (the possibility of) such anomalous cases? If they could not, or not without prohibitively difficult individual research, unfairly bad consequences would surely follow. It seems implausible that linguistic adaptation would eliminate all bad consequences from Y>X mismatches. The weak version of the contingency objection is unpersuasive.

A stronger version of the objection would be that misleading labels can lead to *good* consequences. Perhaps inaccurate, coarse-grained crime labels avoid comparatively worse legislative cost, complexity, and uncertainty.[53] Or perhaps misleading labels result in appropriate caution: perhaps the police would hire *worse* candidates if they had *more* accurate labels. (Perhaps it is better not to hire any 'assaulter', 'minor pushers' included). If we're interested in the consequences of our labelling, we cannot look only at labels that result in unwarrantedly bad consequences for the offender (false positives). We must also look at labels that result in unwarrantedly lenient consequences for the offender (false negatives). Whether the consequences of our labelling are good or bad on net will depend (in part) on the prevalence and magnitude of these respective errors. The harms from false negatives—overly lenient labelling—could be greater.[54] This judgement is reinforced if we prefer the cost of our

[53] Chalmers and Leverick, 'Fair Labelling in Criminal Law' (n 14) 239. There is additionally the risk of leaving lacunae in the law. This can be avoided by retaining a catch-all generic label (like 'assault'), but this replicates (albeit in mitigated form) the grounds of the complaint in *Job Application 2*.

[54] False negatives may result in public dissatisfaction with the official response to crime, sparking (net worse) 'self-help' solutions like vigilantism.

errors to fall on the victims of false positives (convicted offenders) rather than the victims of false negatives (innocent bystanders). Unfairly strict labels may lead to better consequences than even accurate labels.

The problem with the strong objection is that a statement is (objectionably) defamatory if it results in bad consequences *for the defendant*, not if it results in bad consequences *on net*. The (possible) fact that others would be benefitted doesn't make a statement non-defamatory. That would turn the concept of moral defamation into a much broader standard of consequence-evaluation. Whether the net consequences justify the (morally defamatory) bad consequences for the defendant is certainly important. But it's not the issue under contention here.[55] Our objector can still press: what if legislative simplicity, certainty, avoiding false positives, etc all make even mislabelled defendants better off? Two responses. First, this now seems highly implausible. It stretches credulity to imagine that the defendant in *Job Application 2* was nonetheless made better off by their mislabelling.[56] Second, however, I'll simply bite the bullet for the remaining cases. If the category of labelled defendants wouldn't face (net) bad consequences, this is simply not objectionable defamation.[57] We shouldn't object.

2.5 Unreasonableness

It would be easy to object to misleading labels which resulted in the law itself imposing bad consequences, like overly harsh sentences or draconian offender registers. But that easy objection is not to unfair *labelling*, but rather to overly harsh sentences and draconian offender registers.[58] The law could drop those harsh consequences without changing the label. We are only objecting to the bad consequences *of labelling* if those consequences are not (easily) severable from the label. That's true only where the legal system does not control the consequences. That's why the defamation argument focuses on public beliefs. But if the bad consequences are imposed by others, shouldn't the defamation argument be addressed primarily *to those others*? The police in *Job Application 2* extrapolated too much from the label

[55] Section 3. See too Manwaring, 'Criminal Law Principles' (n 2).

[56] Indeed, we care about the consequences for the *category* of defendants labelled (not token defendants), which makes it even less plausible that the consequences will turn out net positive.

[57] Recall again that I don't claim to answer how bad the consequences must be, or for how many defendants, for a crime label to count as morally defamatory.

[58] One exception is that convictions need to be accurately labelled if future sentencing decisions are to rely on them. Chalmers and Leverick, 'Fair Labelling in Criminal Law' (n 14) 231. But better recording of the details of the case would solve this; misleading labels are non-central.

'assaulter', placing too much reliance on a flawed heuristic. Isn't this *their* problem? Yes, it is. But it's *also* the law's problem. The law can't just disown the predictable bad consequences of its actions, especially not if it can mitigate those consequences with better labels. It can't justify throwing defendants to the lions on the basis that the lions shouldn't have eaten them.[59]

But that is a little too quick. The defamation argument indexes the permissibility of crime labelling and conviction to certain public beliefs (Y). Now, some theorists may welcome this indexing. Victor Tadros claims that the criminal law can contribute to 'fostering and sustaining…social solidarity' by tracking the 'reasonable moral convictions of the community'.[60] Going further, Jeremy Horder argues that crime definitions should reflect public beliefs as ascertained through opinion surveys.[61] And, famously, Lord Devlin argued that the criminal law ought to enforce certain strongly held public beliefs.[62] But indexing to public beliefs is usually considered a problem. Unlike Tadros, the defamation argument makes no exception for erroneous or otherwise unreasonable public beliefs.[63] Should these nonetheless translate into policy prescriptions? As Horder makes no such caveat for unreasonable beliefs, Andrew Cornford points out that 'popular judgment [might] attach…significance to factors that the law has traditionally regarded as irrelevant: such as the killer's motive, the factual circumstances of the killing, or the identity of the victim', all considerations that Horder may pause to endorse as relevant.[64] Finally, Lord Devlin's view is highly unpopular: if the public endorse bad norms, shouldn't the criminal law *resist* rather than accommodate those beliefs? It seems like indexing permissible crime labels and convictions to public beliefs is a problem for the defamation argument.

---

[59] As Simester puts it, 'the state may not disregard…its audience, and the effect that such a label will have on D's life.' 'Is Strict Liability Always Wrong?' (n 1) 36.

[60] Tadros, 'Fair Labelling and Social Solidarity' (n 14) 79.

[61] Jeremy Horder, *Homicide and the Politics of Law Reform* (Oxford 2012) ch 1.

[62] Patrick Devlin, *The Enforcement of Morals* (Oxford 1968).

[63] 'Just as equal opportunities policies may alienate sexists and racists, so may stricter laws on domestic abuse, rape, and inciting racial hatred. But this is hardly an argument against stricter laws on domestic abuse, rape, and inciting racial hatred.' Tadros, 'Fair Labelling and Social Solidarity' (n 14) 76.

[64] Andrew Cornford, 'The Architecture of Homicide' (n 24), 837. (Citations omitted). Cornford notes that Horder's 'substantive recommendations…are not supported by the kind of systematic public opinion research that [he] advocates. Indeed…they are based on the very kind of "argument from principle" that the first chapter purports to discourage.' (At 833). He adds: 'Imagine that, if ordinary citizens were to become directly involved in the homicide reform process, the result would be a much worse law of homicide. In that case, it is difficult to believe that this involvement would be inherently valuable or worthy of legislative deference. Even if it were inherently valuable, it is difficult to believe that its value could outweigh the disvalue associated with a much worse law of homicide.' (At 834).

Luckily, the traditional worries about enforcing public beliefs are a red herring here. The defamation argument does not prescribe *enforcing* public beliefs. To the contrary, it prescribes *reacting* to them. This often implies diametrically opposed policies. Consider a public who think littering is heinously culpable. To simplify, Devlin claims that we should enforce those strong anti-littering beliefs via correspondingly serious criminal sanctions.[65] The defamation argument says the opposite: as the public belief in litterers' culpability (say, Y_{50}) dramatically exceeds offenders' actual culpability (say, X_5), the law must instead find some way to change the 'litterer' label to *prevent* the public overreaction. If there is no suitable label, the law must not convict litterers. If it did, it would be complicit in the bad consequences wrought by the public overreaction. Thus, it doesn't matter if the public reaction is unreasonable: the defamation argument's conclusion is calibrated precisely to dampen such responses.[66]

But that is an easy case. The public might be too *vindictive* (Y>X) or too *permissive* (Y<X) in ways that render the defamation argument either too strong or too weak.[67] *Vindictive* public beliefs extend to serious crimes: imagine the public believe moderately culpable thefts (X_{30}) are atrociously culpable (Y_{100}).[68] It may be possible to relabel 'theft' to avoid the overreaction. But, failing that, if we can't avoid the Y>X discrepancy, the defamation argument bars the conviction of *bona fide* culpable thieves. That seems too strong. *Permissive* public beliefs present the opposite problem. Imagine the law labelled very minor sexual assault (X_{25}) as 'rape'. This seems like an obvious case of defamation. But if the public is unreasonably permissive

[65] Devlin's prescription is conditional on certain harmful outcomes occurring given a failure to enforce public beliefs.

[66] Simester imagines a society that treats parking offenders like paedophiles, and notes '[e]ven in this sort of case, at least where the stigma is predictable,…the state should take account of the consequences of a conviction for defendants.' 'Is Strict Liability Always Wrong?' (n 1) 36. Chalmers and Leverick discuss unreasonable public reactions in the context of media bias in 'Fair Labelling in Criminal Law' (n 14) 228-229. Moral philosophers split as to whether or how much we ought to predicate our own (evaluation of) actions on the actions of others: Bernard Williams opposes such conditionalization in Bernard Williams and JJC Smart, *Utilitarianism: For and Against* (Cambridge 1973), while Jeff McMahan asserts that 'individuals must decide what to do against the background of what others will in fact do'. Jeff McMahan, 'Philosophical Critiques of Effective Altruism' (2016) 73 The Philosopher's Magazine 92. In the context of the criminal law, see John Gardner, 'Complicity and Causality' in *Offences and Defences* (n 15).

[67] The public may also (1) be *accurate* (Y = X), or (2) simply have *no relevant view* (no Y exists). If *accurate* (public beliefs about culpability track actual culpability) the defamation argument has no objection. If there's *no belief* (eg for white collar or otherwise obscure crimes), this results in the same problem as too permissive beliefs (discussed below), meaning the law could fail to take (say) accounting fraud seriously enough. But ignorance is amenable to relabelling: *some* label will convey appropriate censure (eg 'corporate theft'). Neither therefore present a problem.

[68] Consider righteous social media vigilante mobs, catalogued in Jon Ronson, *So You've Been Publicly Shamed* (Picador 2015).

regarding rape—say, Y_{25}—then it is not the case that Y>X, and thus the defamation argument offers no objection. As it seems like it *should* object—it fails to take rape seriously enough—the defamation argument appears too weak.

The problem of permissive audiences is avoidable. We have reasons not to use harsh labels even if a permissive public fails to take them seriously. This could be on entirely separate grounds from the defamation argument. But we could also interpret Y to include *reasonable* subsets of the public.[69] That is, we can class statements as defamatory if *either* Y^{ACTUAL} *or* $Y^{REASONABLE}$ exceed X. This brings the argument closer into line with the tort of defamation, which indexes defamation to reasonable beliefs.[70] At first this seems quite distinct from indexing to actual beliefs: surely bad consequences require actual, not hypothetical, audience reactions? But, regardless of how tort *identifies* defamation, its primary *remedy* is to award damages to compensate actual injury suffered, both reputational and monetary.[71] Thus tort's 'reasonable' audience can be thought of as a subset of the public whose beliefs materially affect the victim. This is an additional specification of the relevant 'Y'.[72] Permissive public beliefs do not present a problem for the defamation argument, then, so long as Y can be composed of some reasonable subset of the public. If *they* appropriately judge rape as Y_{100}, then labelling minor sexual assaults (X_{25}) as 'rape' will indeed constitute defamation.

---

[69] See fn 26, above.

[70] The classic test is from *Sim v Stretch* [1936] 2 All E.R. 1237, 1240: '[W]ould the words tend to lower the plaintiff in the estimation of *right-thinking* members of society generally?' (Lord Atkin). The qualification '*reasonable* people generally' was used by the Home Office, ['Faulks'] Committee Report on Defamation Cmnd. 5909 (1975), cited in *Berkoff v Burchill* [1997] E.M.L.R. 139, 144. Similarly, when asking whether a statement was one of opinion (versus fact, required to benefit from the defence of honest opinion), the court asks 'how it would be interpreted from the perspective of the ordinary *reasonable* reader'. Alastair Mullis and Andrew Scott, 'Tilting at Windmills: the Defamation Act 2013' (2014) 78 MLR 87, 92 (emphases added).

[71] '[I]t is the award of damages, not the grant of an injunction…which is the primary remedy which the law grants on proof of this tort'. *John v MGN Ltd* [1997] QB 586 at 607 (Lord Bingham). The cost of injury depends on the injured. The rich and famous have more to lose, and thus potentially higher damages.

[72] In the other direction, tort has sometimes compensated reputational damage even for highly unreasonable beliefs. When Princess Natasha complained of an innuendo that she slept with Rasputin, Slesser LJ stated that he 'cannot see that from the plaintiff's point of view it matters in the least whether this libel suggests that she has been seduced or ravished.' It would be highly unreasonable to disregard the distinction between seduction and rape, but reputational damage occurred regardless, and the law sought to rectify it. *Youssoupoff v MGM Pictures Ltd* (1934) 50 TLR 581.

The problem of vindictive audiences is trickier. It's also more pressing: most publics are likely too vindictive in general.[73] I don't think the defamation argument has any simple solution. Trying to prevent public overreactions is very difficult. In many cases, the only feasible way to avoid overreactions would be to avoid convictions, even for culpable crimes. That seems wrong, and thus the objection seems right. But this doesn't mean we should abandon the defamation argument. Rather, we should accept that its conclusions are not absolute. The reasons to convict moderately culpable defendants may outweigh the reasons not to defame them. That is, the law should convict such defendants even if the public will overreact to that labelling. This doesn't entail indifference to that outcome. If the law cannot comply with its primary duty not to defame, it could yet do the next best thing. It could fulfil a secondary duty to improve public beliefs to avoid that outcome.[74] In time, improved public beliefs will mean criminal labels and convictions no longer constitute defamation amongst the (newly enlightened) public. The law could avoid the problem of unreasonable audiences by *making* audiences reasonable. Of course, there are no guarantees of reliable or significant public belief improvements. My claim is just that the force of the defamation argument doesn't disappear even if outweighed. Instead, the argument's prescriptions flow through to imply secondary duties.[75]

[73] Most publics are largely ignorant about the content and workings of criminal law. They tend to think that crime is rising regardless of reality, that sentencing is much more lenient than it is, and that sentencing ought to be more punitive. Michael Hough and Julian Roberts, 'Attitudes to punishment: findings from the British Crime Survey' (1998) Home office research studies; Francis Cullen, Bonnie Fisher, and Brandon Applegate, 'Public Opinion about Punishment and Corrections' (2000) 27 Crime & Just 1; Julian Roberts et al, *Penal Populism and Public Opinion* (Oxford 2003) ch 2. The US General Social Survey dataset offers some insight: over 10% of respondents thought the harshest available punishment options (10-20 years or life imprisonment) were appropriate for government/military employees who commit minor crimes like leaking information, downloading pornography to government computers, or stealing truck parts. Data available at http://sda.berkeley.edu/sdaweb/analysis/?dataset=gss14nw (cumulative 1972-2014 with codes PUNLEAK, COMPORN, and PUNTRCK). Canadians have similarly vindictive views, while Germans are only slightly less punitive: Matthew Kugler et al, 'Differences in Punitiveness Across Three Cultures: A Test of American Exceptionalism in Justice Attitudes' (2013) 103 JCrimLaw and Crimlgy 1071.

[74] Leslie Green, 'Should Law Improve Morality?' (2013) 7 Criminal Law and Philosophy 473. Even Jeremy Horder endorses a (qualified) educative role for criminal law academics: Horder, *Homicide and the Politics of Law Reform* (n 61) 29-32.

[75] This thought is familiar from John Gardner's 'continuity thesis' in 'What is Tort Law For? Part 1. The Place of Corrective Justice' (2011) 30 Law and Philosophy 1. Ashworth suggests some steps the law could take in Ashworth, 'Ignorance of the Criminal Law' (n 1).

This response raises a final difficulty: if the defamation argument's proscription of moral defamation is not absolute, how strong is it? Without knowing the rough strength of a culpability principle, it is impossible to use it to evaluate the criminal law.

3 Strength

Recall the defamation argument's conclusion:

> (6) …it is morally objectionable to use crime labels where the public belief in the degree of culpability which attaches to that label exceeds offenders' actual culpability

How morally objectionable? Andrew Simester claims that it is *impermissible*. He says that 'the right not to be wrongly censured [morally defamed] defeats *any* instrumental case for strict liability in stigmatic crimes.'[76] As we noted above, his argument generalises to other contexts of conviction disproportionate to culpability. On Simester's view our culpability principle is an absolute constraint.

But absolute constraints are rarely plausible. No matter how wrong we believe moral defamation to be, justificatory circumstances are easy to conjure. Philosophers often rely on extreme hypotheticals: it's justified to push someone in front of a train *to save the planet*. Any allegedly absolute constraint on killing will turn out to have limits.[77] The same goes for moral defamation. I have already suggested some limit cases: for moderately culpable crimes, even if Y>X it may be better to defame than decriminalise. The criminal law is used to these trade-offs. Punishing the innocent warrants a strong constraint if anything does. To avoid doing so, the criminal law requires proof of guilt beyond a reasonable doubt.[78] Even this results in

[76] Simester, 'Is Strict Liability Always Wrong?' (n 1) 37.

[77] Larry Alexander and Michael Moore, 'Deontological Ethics' in Edward Zalta (ed), *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition) §4 for discussion of 'moral catastrophe' cases.

[78] Usual glosses on the probability threshold required for conviction beyond a reasonable doubt (95%, 99%, etc.) assume that juries are well calibrated, ie that their subjective probability, the one they decide cases based on, matches reality. This is improbable: most people are overconfident. Hence the *de facto* standard may be lower than 95%, perhaps much lower. Nor will this effect being cancelled out by multiple jurors, assuming they tend to be overconfident in the same direction.

innocent people being punished.[79] An absolute constraint would require *certainty* or abolition.[80] Neither option is plausible, including to Simester. He accepts that 'instrumental considerations…permit wrongful convictions whenever the criminal proof standard is met.'[81] So why does he endorse an absolute constraint on defamation? He explains:

> Where guilt is proved beyond reasonable doubt in stigmatic crimes, the state convicts in *good faith*—D is believed to be culpable. Further, although error is systemic it remains *unsystematic*: the distribution of error is unknown, and we cannot predict the likelihood that any particular conviction is a mistake. By contrast, where strict liability is employed in a stigmatic crime, the state consistently labels D as a culpable wrongdoer without believing this to be true. Moreover, defamation is predictable—there are reasons for thinking that the state is particularly likely to censure and punish D wrongly in that class of cases. Hence, while instrumental considerations of an institutional nature may sometimes be relied upon to justify the risk of good-faith erroneous convictions, arguments of this type seem inadequate to justify strict liability for stigmatic crimes.[82]

This doesn't vindicate the plausibility of absolute constraints. We can always think up extreme hypotheticals in which, all things considered, we should *systematically* convict in *bad faith*, ie those believed to be non-culpable. The constraint against moral defamation might be weighty, but not absolute. Simester's argument is useful, however, in suggesting that systematicity and bad faith are factors which should be given significant *weight* when judging the strength of the constraint.

---

[79] 1.23 million defendants were convicted in England and Wales from April 2016 to March 2017. If only 1% of convictions were erroneous, that is 12,300 defendants. See Ministry of Justice, 'Criminal Justice Statistics quarterly, England and Wales, April 2016 to March 2017 (provisional)', 17 August 2017. (Of course, most defendants plead guilty. But we can assume that a guilty plea is a partial function of the probability of being punished, including erroneous punishment, and thus (or for other reasons) will include some innocent defendants pleading guilty.

[80] This point is made in another context by Victor Tadros, *Wrongs and Crimes* (Oxford 2017) 95. Perhaps we may justify a categorical constraint of an *actor* (eg the state when convicting) absent a *categorical moral norm*, as an error-reduction mechanism. See Les Green, 'The Nature of Limited Government' in John Keown and Robert George (eds), *Reason, Morality and Law* (OUP 2013), discussed in James Edwards, 'Master Principles of Criminalisation' (2016) 7 Jurisprudence 138, 146, and Manwaring, 'Criminal Law Principles' (n 2). That is, morality and axiology may come apart. But the reason to convict the culpable may be sufficiently strong to outweigh (or defeat) the reason not to do so *even accounting for the meta-reason in favour of absolute constraints*, just by ratcheting up the hypothetical consequences. The strength question remains.

[81] Simester, 'Is Strict Liability Always Wrong?' (n 1) 36.

[82] ibid 36-37.

However, these (alleged) features of strict liability cases don't necessarily generalise to other crime labels disproportionate to culpability. Defamatory crime labels need not be systematic, and lawmakers may not have any beliefs as to the public belief (Y). If *most* defamation cases are non-systematic or in good faith, this might imply a very *weak* constraint on defamation, easily overcome by instrumental considerations. That would undercut the practical significance of our culpability principle. Three responses. First, I doubt that systematicity has the significance that Simester suggests. Second, I doubt that certain types of defamation are in good faith. Third, these two considerations don't give us the relevant baseline strength of the culpability principle.

Take systematicity first. Simester claims that it is permissible to convict the non-culpable due to non-certain proof standards as the distribution of error is unknown ('systemic') but impermissible if the distribution of error is known ('systematic'), which he claims applies to convictions for strict liability. Two problems. First, this description is misleading. Strict liability is often used to target genuinely culpable conduct where proving intention or recklessness as to that conduct (or a sub-element) would be prohibitively difficult. Terrorism offences are the obvious example. The distribution of 'error' as to culpability in strict liability crimes is unknown, just as with proof standards.[83] Conversely, the error distribution might be reasonably well-known for proof standards: presumably some factors (like good lawyers) will skew the error rate down (at least for erroneous convictions), while certain crimes may have predictably higher error rates (like those which rely exclusively on testimonial evidence). Second, it's not clear why knowledge of the error distribution makes this normative difference. Some crimes, and especially some sub-elements, have predictably higher error rates due to the difficulty of ascertaining their truth. Any proof of intention will be less certain than proof of negligence as to the same conduct, simply due to the difficulty of inferring mental states. This may be one reason not to require proof of intention, but a reason of very little weight. That reason will be dwarfed by other relevant considerations, like requiring high culpability. Likewise, that strict liability (might) result in higher error rates (as to culpability) doesn't count

[83] True, the law doesn't *say* that non-culpable convictions are erroneous if there is a strict liability standard. But culpability is often *de facto* necessary for officials to exercise their discretion to arrest or prosecute (eg because it affects the public interest). Further, courts take a very restrictive view of strict liability standards (eg re: conclusive presumptions of non-consent in sexual offences per the Sexual Offences Act 2003 s76: *Jheeta* [2007] EWCA Crim 1699, *Bingham* [2013] EWCA Crim 823). Analogously, the law doesn't *say* that mistaken convictions where there is proof beyond a reasonable are legal errors. But such decisions can nonetheless be overturned with new evidence, as could (at an earlier stage) the decision to prosecute non-culpable strict liability offences. These are differences in degree rather than kind.

as a strong reason against using it. Other considerations will dwarf it in any cost-benefit analysis.

The case for giving weight to bad faith is stronger. Simester's gloss is that convicting D in the belief that D is not culpable shows bad faith. The point isn't that officials believe in the *possibility* of error, or even that a certain class of defendants attracts *more* errors. Those errors may be unavoidable and justifiable.[84] Rather, it is especially objectionable to convict believing that *this* class of convictions *are* erroneous. This shows bad faith because known errors are avoidable.[85] It is worse for a judge to ignore exonerating evidence than to be merely ignorant of it.[86] It is worse for the criminal law knowingly to convict disproportionately than to merely risk such convictions. *Obvious* Y>X mismatches will tend to be known, avoidable, and therefore in bad faith. Our culpability principle will be especially stringent in these cases. It is very hard to justify convictions of the most serious crime labels—murder, rape, etc—without correspondingly high defendant culpability.[87]

Even if we know that bad faith affects the weight of our culpability principle, we still don't know its relevant baseline strength. When can its normative demands be overcome? Will that require very strong or only mild countervailing instrumental considerations? I can't offer a complete answer. To begin with, it is important to acknowledge that it requires overcoming at all. I find it plausible that justifying derogations from the culpability principle would take weighty countervailing considerations.[88] But this thesis is about incapacity rules, and it is here that the culpability principle speaks most clearly. My final claim is that it rules out a common argument used to justify convicting incapably non-culpable defendants.

Criminal lawyers discussing (prosecuting, judging) incapacity rules often run two arguments at the same time. First, they argue that a defendant was culpable *despite* their incapacity.

---

[84] Poorer defendants may predictably attract more errors than wealthier defendants given unequal access to legal advice. But this may be unavoidable if there is no feasible/justifiable way to bridge that gap. (It could remain a problem for political morality).

[85] There are exceptions (eg n 86), but these would rarely apply to the argument above.

[86] Excepting inadmissible evidence, regarding which judges must act *as if* ignorant. That is to be justified by broader systemic considerations.

[87] Defamation being obvious also strengthens the secondary duty to improve public beliefs by removing a reason against imposing that duty; namely that officials may be ignorant of it.

[88] I leave open whether the constraint is a weighty reason or an exclusionary reason (of whatever scope), and what it would take to outweigh/overrule/cancel or otherwise defeat it. Joseph Raz, *Practical Reason and Norms* (2nd edn, Oxford 1990); John Gardner, 'Justifications and Reasons' in *Offences and Defences* (n 15).

Demonstrating culpability helps to justify conviction, a nod to an implicit culpability principle. Second, they argue that even if the defendant was non-culpable, their incapacity makes it appropriate to convict for policy reasons of public protection. This argument abandons any implicit culpability principle. These are very different arguments. They require very different justifications. Yet criminal lawyers sometimes move freely between the two without acknowledging the necessary shift between justificatory modes.[89] My claim is that the second argument almost never suffices to overcome the strength of the culpability principle.

Consider sexual offences. A necessary condition of several sexual offences is that D does not *reasonably believe* that V consents.[90] They do not require that defendants actually advert to V's non-consent. In this sense these offences are objective. Some defendants may have certain incapacities that materially affect their beliefs in ways that render those beliefs objectively unreasonable. A schizophrenic D may deludedly believe in V's consent. An autistic D may fail to understand V's sarcastic 'consent'. These defendants, believing V consents, taking what they believe to be all reasonable precautions, may initiate sexual contact. In some cases, their incapacities could make it the case that these individuals are not culpable for their objectively unreasonable beliefs and subsequent conduct.[91] Our culpability principle would count against labelling and convicting these defendants as 'sexual assaulters', as these labels imply very serious culpability. But they can be convicted and labelled as sexual assaulters nonetheless. By not requiring advertence the criminal law does not ensure culpability. It risks morally defaming these individuals. And it does so fully advertently.

I do not claim that our culpability principle should rule out any response to these incapable defendants, or even any criminal law response. I'm not well placed to argue, all things considered, whether it's justified, on policy grounds, to control certain people. I claim only that this control shouldn't occur in the name of criminal culpability. That is, it shouldn't occur

[89] Eg *Martin* [2001] EWCA 2245 [65]-[67] noted *Smith (Morgan)* [2001] 1 AC 146's claim that 'it would be unjust not to take into account' certain characteristics of the defendant in determining what counted as loss of control. The court in *Martin* refused to follow that test regarding self-defence for various 'policy reasons', including increased medical disputes. The court did not try to justify the differential approach in terms of justice or D's culpability. Similarly, in *C (Sean Peter)* [2001] EWCA Crim 1251 [11]-[24], counsel argued that fairness demanded that harassment's negligence standard should account for D's mental illness (presumably as such Ds are thought non-culpable). The court rejected that argument entirely by reference to the relevant Act's 'protective and preventative' purpose rather than by meeting the challenge of unfairness due to non-culpability. (See ch 3 §3.2 and fn 173).

[90] Eg Sexual Offences Act 2003 s1.

[91] Obviously other conditions must be absent or present: most cases will involve circumstances that renders this sort of defendant culpable (eg other obvious warnings, etc.) I claim only that there may be *some possible* cases of non-culpable defendants who hold an unreasonable belief as to consent.

via conviction, or at least via conviction under such serious labels. To do so constitutes advertent moral defamation of incapably non-culpable defendants. This could yet be justified in preference to decriminalisation. But that is not the only alternative. The law has other options. It has the special verdict. Insane defendants are frequently dangerous and justifiably subjected to coercive measures as a matter of public policy. But this control occurs without conviction of a stigmatic crime: defendants are *not guilty* by reason of insanity. Public protection orders do not require conviction.[92] The same model could apply to other incapable defendants. However strong our culpability principle may be, it must rule out the advertent conviction of non-culpable defendants where it is entirely possible to avoid doing so, without sacrificing any public protection. Of course, culpability is a matter of degree. I'm not claiming that *any* exculpatory incapacity presents an insuperable bar to *any* criminal convictions whatsoever if there happens to be alternatives available. One of the alternatives to conviction for very serious offences might be conviction of a less serious offence.[93] I claim only that the law ought to utilise such alternatives. It should account for defendants who are morally exculpated on the basis of incapacity by providing some form of legal exculpation the better to tailor Y to X.

This is one practical upshot of my culpability principle. As we will see in the next chapter, the law contains many rules to avoid such moral defamation. But the culpability principle remains frequently violated.

[92] Existing sentencing possibilities include non-custodial sentences for almost all crimes: see *B(MA)* [2013] EWCA Crim 3 at [40]. My proposal simply shifts the precursor from conviction to a kind of special verdict.

[93] This, indeed, is how the partial defences to murder work: they reduce a murder conviction to one of manslaughter.

2 Incapacity Rules

Our central question is: how should the criminal law account for incapacities? It would help to know first how the law *does* so. Call rules which account for incapacities *incapacity rules*. This chapter has two goals: to produce a taxonomy of incapacity rules (§1), and to populate it (§2-4).[1] We'll see that incapacity rules are more numerous and more varied than is usually thought.

1 Three types of incapacity rule

1.1 Incapacity doctrines

Insanity and infancy are often said to be the only rules that account for incapacities, because doing so is their entire point. Call them:

> **Incapacity doctrines**: legal rules whose rationale is to account for incapacities.[2]

Incapacity doctrines are the first type of incapacity rule.

Why look for rationales? Because identifying the main reasons for a rule is the best way to understand them. The reasons to prevent (and/or punish) assaults and fraud explain why we have offences of assault and fraud. We simply don't understand these offences if we don't understand those rationales. The same cannot be said for other aspects of rules. Insanity and

[1] You might doubt the methodological proprietary of studying a non-legal category of rules. If so, consider an analogy. A sexual dimension to wrongdoing makes a normative difference that should affect the shape of the criminal law. It would be useful to study 'sexual offences' even if the law recognised no such category. Likewise, I think, for incapacities. The modern law of restitution provides a clear example of the power of imposing a taxonomy on (apparently) diverse legal doctrines. Peter Birks, 'The Concept of a Civil Wrong' in David Owen (ed), *Philosophical Foundations of Tort Law* (Oxford 1995).

[2] A person may be *legally* incapable of committing an offence, eg if they have diplomatic immunity. This is a rule's *output* (eg n 36) But we're interested in incapacities which form the rule's *input*. Sometimes this distinction is elided. I call any rule or bundle of rules that contains a distinct rationale a 'doctrine'. I take no view on their individuation: see Joseph Raz, 'Legal Principles and the Limits of Law' (1972) 8 Yale LJ 823, 823-829.

infancy *appear* quite different. Insanity is defined as a 'defence', by common law, with a multi-pronged and vague test. Infancy is defined as a 'presumption against an offence', by statute, with a short and precise test. But these differences are less important than their shared rationale. That's why textbooks treat them together.

How do we find rationales? Primarily by interpreting legal authorities (the standard doctrinal method), sometimes supplemented with preparatory, comparative or historical sources. The rationales so uncovered may be explicit or implicit. That doesn't matter. What matters is finding the law's main reasons for having the rule.

How do we know if a rule's rationale is to account for incapacities? One approach would be to define what counts as an incapacity, to look for rules with the relevantly defined features, then conclude that those rules are incapacity doctrines. But one could offer a biased definition of incapacities which would skew the resulting analysis.[3] Instead, I'll use authorities' own accounts of incapacities. Incapacity-oriented language would help, but the question is ultimately interpretive. We're not asking *which* incapacities count, only *whether* a rule's rationale is to account for incapacities.[4]

I'll argue that, along with insanity and infancy, fitness to plead, automatism, diminished responsibility, and loss of control may all count as incapacity doctrines.

1.2 Relativisations

There are many legal rules whose rationale is not to account for incapacities, but which have incapacity-based elements. Consider reasonableness. Criminal law is suffused with reasonableness standards. What counts as reasonable is often relativised to some of the defendants' personal attributes. These are:

Relativisations: standards relativised to defendants' attributes

The law could relativise to any attribute. It usually selects age and sex. For example, the loss of control defence to murder requires that defendants demonstrate 'a normal degree of

[3] We might (consciously or otherwise) define incapacities such that whichever doctrines we think *should* account for incapacities *would* account for incapacities per our (biased) definition.

[4] See ch 5.

tolerance and self-restraint' for 'a person of D's sex and age…and in the circumstances of D'.[5] It seems like the standard of tolerance required is relativised to D's age, sex, and circumstances. An old man would be judged by the calmness of seniority; a police officer to her employment circumstances. Relativisations are indexed to various attributes of a reference class. The idea is that what we're entitled to expect from a defendant varies with these attributes. We're interested in relativisations to incapacities, or:

Incapacity relativisations: standards relativised to defendants' incapacities

Given the ubiquity of (reasonableness) standards, incapacity relativisations potentially affect far more defendants than the incapacity doctrines.[6] But there are two difficulties with identifying incapacity relativisations.[7]

The first difficulty affects all relativisations. The relevant attributes (including incapacities) may affect liability without constituting relativisations. Take loss of control again. A provocation only counts as a 'qualifying trigger' if it is 'extremely grave'.[8] This is a fixed standard. We're entitled to expect everyone to resist provocations that are not so grave. Now consider the provocation 'You should kill yourself.' This is hurtful but probably not 'extremely grave'. Unless, that is, it is directed at someone suffering suicidal depression. D's depression, an incapacity, makes it the case that the provocation is extremely grave and thus that the defence is available.[9] D's incapacity affects her liability. But this is not a relativisation. Her incapacities do not affect what we are entitled to expect from her; they do not relativise the standard applied. Rather, her incapacities are facts which bear on whether a *fixed* standard of 'extreme gravity' is met. Similarly, D's ethnicity may affect whether a racial slur is 'extremely grave'. But this does not mean the standard of 'extreme gravity' is relativised to ethnicity. D's attributes only *incidentally* affect her liability. They are:

[5] Coroners and Justice Act 2009 (CJA) s54(1)(c).

[6] While incapacity doctrines technically contain incapacity relativisations, we'll reserve the latter term for elements of doctrines which have *other* primary rationales.

[7] Additionally, some claim that relativising reasonableness standards to account for incapacities is conceptually confused. I discuss this and other critiques in ch 3.

[8] CJA ss54(1)(b), 55(4)(a).

[9] Assuming the other elements are made out. *Wilcocks* [2016] EWCA Crim 2043 and *Rejmanski* [2017] EWCA Crim 2061 both involved taunts about D's suicidal depression.

> **Incidentally relevant attributes**: attributes which factually affect whether a standard applies but not by altering the standard itself

Above I said it seems like loss of control's standard of tolerance is relativised to D's age, sex, and circumstances. But perhaps age, sex, and circumstances *incidentally* affect whether a *fixed* standard of tolerance is met, just as suicidal depression and ethnicity may incidentally affect whether a provocation counts as 'extremely grave'. Is that what the section's reference to 'a *normal* degree of tolerance and self-restraint' intended? If so, loss of control's standard of tolerance is *not* relativised. The correct answer seems to be that the 'normal degree of tolerance' standard *is* relativised to age and sex, while 'the circumstances of D' are merely incidentally relevant.[10] It would be easy to misunderstand the law if we didn't distinguish relativisations from incidentally relevant attributes.

The second difficulty affects only incapacity relativisations. The law rarely relativises *explicitly* to incapacities as it (seemingly) does to age and sex.[11] This doesn't mean it *isn't* doing so. Perhaps age and sex include corollary incapacities. If so, loss of control's standard of tolerance is *implicitly* relativised to D's capacities:

> **Implicit incapacity relativisations**: standards relativised to defendants' incapacities via relativisations to other attributes

Given the scarcity of explicit incapacity relativisations, we'll have to identify implicit corollary incapacities within relativisations to identify incapacity relativisations.

In sum, to identify incapacity relativisations we must ask two questions. First: is a standard relativised? (As opposed to invoking merely incidentally relevant attributes.) Second: if so, does

---

[10] *Rejmanski* (n 9) [25]-[28]: as to whether a mental disorder counted as a relevant 'circumstance', the court held that it 'must not be relied upon to undermine the principle that the conduct of the defendant is to be judged against "normal" standards, rather than the abnormal standard of an individual defendant.' Given that there is no 'normal' sex or age, presumably this normality requirement does not extend to those attributes and thus the section *does* relativise to them. (Reinforced by the section's precise formulation). *Rejmanski* did leave open the possibility of 'other circumstances where a disorder might be relevant' to the standard, but added that it was 'of academic interest only.' One (academic) suggestion: disorders affecting D's mental *age*, or gender dysphoria affecting D's (perceived) *sex* may be relevant (and relativised) by analogy.

[11] One exception was the (now abolished) doli incapax defence, which asked whether D was less *capable* than his peers.

the relativisation account for incapacities (including implicitly)? If so, it is an incapacity relativisation.

I'll argue that the standards of reasonable firmness in duress, reasonable care in negligence, and reasonable belief in sexual offences all incorporate incapacity relativisations.

1.3 Counterfactual relativisations

Many doctrines (including those just mentioned—duress, negligence, and sexual offences) use a two-prong test: (1) What did D believe? (2) Given that belief, did D act reasonably? Sometimes the first prong sets a standard: what *should* D *reasonably* have believed? If it does, and if this standard is relativised to D's incapacities, then the first prong constitutes an incapacity relativisation. But sometimes the first prong sets no standard: it just asks what D *genuinely* believed. If it doesn't set a standard, a fortiori it doesn't set a relativised standard. It can't constitute an incapacity relativisation. It can't be an incapacity rule.

But the last sentence is mistaken. The law might ask what D genuinely believed *rather than* what they reasonably should have believed precisely to account for (belief-affecting) incapacities. The rationale for *not* setting a standard might be to account for incapacities. Call these:

> **Counterfactual relativisations**: legal rules whose rationale for the absence of a standard (which could plausibly have been imposed) is to account for defendants' attributes

They include:

> **Counterfactual incapacity relativisations**: legal rules whose rationale for the absence of a standard (which could plausibly have been imposed) is to account for defendants' incapacities.

There are three questions we should ask of counterfactual incapacity relativisations. (1) Are they incapacity rules? (2) Do we need to include them in our taxonomy? (3) Are they distinguishable from incidentally relevant incapacities?

First question: are they incapacity rules? Consider a lightly fictionalised legal fable:

> It was an offence for D to engage in non-consensual sexual activity where D did not 'reasonably believe' the other party consented. This standard of reasonable belief was

not relativised. The reasonable person shared none of the defendants' attributes. In due course, cases arose involving defendants with various incapacities. A boy, in youthful inexperience, mistook his girlfriend's comments for encouragement. He sexually touched her. He was charged with sexual assault. The judges were troubled. The boy's belief wasn't reasonable for an adult. But it seemed unfair and unjust to hold the boy's beliefs to that standard. They found him not guilty. They held that his belief in the partner's consent was not unreasonable *for a child of his age*. They relativised the standard of reasonable belief to age. But the hard cases didn't stop. At a bar, a man sarcastically told a woman he wanted her to touch him. The woman was moderately autistic. Misreading the man's sarcastic encouragement, she sexually touched him. She was charged with sexual assault. The judges reasoned that if it's unfair to hold children to the standard of adults, then it's also unfair to hold autistic defendants to the standard of neurotypical defendants. They found her not guilty by relativising the standard of reasonable belief to account for mental conditions. But the hard cases didn't stop. Many defendants offered many incapacities to explain their many (would be) unreasonable beliefs. The Supreme Court took a stand. It was no use relativising the standard to this, that, and the other incapacity. The problem was the standard itself. So the justices did away with it. Henceforth, defendants would not be guilty if they *genuinely* believed the other party consented.[12]

These rules started with a non-relativised standard. Incapacities were irrelevant. They moved to an age-relativised standard, including corollary incapacities. Then they relativised further to make other attributes (and incapacities) relevant. Finally, they made *all* attributes and incapacities relevant by omitting the standard of belief required altogether. It would be odd to describe this evolution as a shift from irrelevant to relevant to merely incidentally relevant incapacities. Better to say that D's incapacities were irrelevant, then relevant, then *fully accounted for* as determinants of belief. That best explains the rules' evolution. The rationale for omitting the standard was to account for incapacities. Counterfactual relativisations combine the rationale-oriented dimension of incapacity doctrines with the relativising dimension of relativisations. They should be considered full members of the incapacity rules club.

[12] Based on *Caldwell* [1982] AC 341; *Elliott v C* [1983] 1 WLR 939; *G* [2003] 1 AC 1034; *B(MA)* [2013] EWCA Crim 3; *DPP v Morgan* [1976] AC 182.

Second question: do we need to include counterfactual relativisations in our taxonomy of incapacity rules? Aside from simply being comprehensive, including counterfactual relativisations avoids a problem of arbitrariness. Let's continue our fable:

> The Supreme Court were satisfied that doing away with the standard of belief for sexual offences was the right thing to do. But more hard cases arose. A psychotic man believed that he was irresistible; that everyone consented to his advances. He sexually touched various people without their consent. Given the Supreme Court's ruling, and clear evidence of the reality of his condition, all agreed that the man lacked the relevant mens rea for the offence. But few were comfortable letting him go free. In response, Parliament created a new Act to tweak the test. Now defendants would not be guilty if they genuinely believed the other party consented, *unless that belief was delusional.*

The Parliamentary tweak creates problems for a taxonomy that excludes counterfactual relativisations. On one interpretation, there is no standard of belief. The defendant is to be judged according to their genuine beliefs, but deluded beliefs are an exceptional case. With no standard, there can be no relativisation. But on another interpretation, the tweak means the law sets a standard of *non-deluded belief*. This standard is *highly* relativised: it accounts for *every* attribute but for delusions. Picking between these interpretations is an arbitrary classificatory choice.[13] If we didn't consider counterfactual relativisations, that arbitrary choice could significantly alter how we look at the tweaked provision. We'd either ignore the rule entirely (as a non-standard), or we'd take it as an extremely strong relativisation. Including counterfactual relativisations within our taxonomy of incapacity rules avoids putting that kind of weight on arbitrary classifications: we include the tweaked provision either way.[14]

Third question: are counterfactual incapacity relativisations distinguishable from incidentally relevant incapacities? Here's the worry. No matter which non-standard we're interested in, the law always *could have* imposed a standard. But then counterfactual relativisations would be

[13] If you think the first interpretation is obviously correct, simply add more exceptions until it becomes a borderline case.

[14] The mere acknowledgement of the (incidental) relevance of certain attributes (like D's circumstances) is arguably a (weak) means of accounting for those attributes. Not *deeming away* the relevance of certain attributes is arguably a (weak) means of counterfactual accounting. These weak acknowledgements really are ubiquitous, so we won't consider them further. But it reminds us that the law has a wide spectrum of means by which to account for attributes. (For examples, the law deems away evidence of victims' sexual history (Youth Justice and Criminal Evidence Act 1999 s41), and of factual consent under deceptive conditions (Sexual Offences Act 2003 s76.)

ubiquitous, and consequently of no interest at all. But this is not a problem. Recall that we defined counterfactual relativisations as rules whose *rationale* for the absence of a standard was to account for incapacities, and where a standard *could plausibly have been imposed*. These conditions are certainly not met for all non-standards. They both require evidence. The best evidence that a standard could plausibly have been imposed is if the law formerly *did* so.[15] The best evidence of the rationale for the absence is the justifications offered when omitting them.[16]

I'll argue that tests of belief in a risk (recklessness), belief in a threat (self-defence), fear of violence (loss of control), and belief in various circumstances (dishonesty offences), all constitute counterfactual incapacity relativisations.

2 Incapacity doctrines

Recall that incapacity doctrines are legal rules whose rationale is to account for incapacities. They are identified by interpreting the legal rationale for the rule.

2.1 Insanity

Insanity provides a complete defence to some mentally disordered defendants. The canonical formulation of the doctrine is from 1843's M'Naghten Rules:

> [T]o establish a defence on the ground of insanity, it must be clearly proved that, at the time of the committing of the act, the party accused was labouring under such a defect of reason, from disease of the mind, as not to know the nature and quality of the act he was doing, or, if he did know it, that he did not know he was doing what was wrong.[17]

[15] Failing that, evidence of doctrinal analogues, serious reform suggestions, or comparative examples may suffice.

[16] If multiple rationales are offered, we must weigh their respective importance.

[17] *M'Naghten's case* (1843) 10 Cl & Fin 200, 210.

The Rules don't refer to incapacities or inabilities. But we could gloss them to ask whether D had an *incapacity* of reason, rendering him *incapable* of knowing the nature of the act or that it was wrong.[18] If this gloss is accurate, insanity is an incapacity doctrine. Is it?

The gloss is supported by a broader reading of *M'Naghten's case*. The Rules derive from Tindal CJ's response to five questions from the House of Lords. Those questions followed Daniel M'Naghten's original trial, where Tindal CJ directed the jury:

> [I]f on balancing the evidence in your minds you think the prisoner *capable* of distinguishing between right and wrong, then he was…liable.[19]

Tindal CJ's later formulation presumably meant to specify the same incapacity-based test. Similarly, Maude J's answer to the Lords asked if the defendant was '*incapable* of knowing right from wrong'.[20] This strongly implies that 'Did not know he was doing what was wrong' means '*Incapable* of knowing he was doing what was wrong'.[21] Similarly, the US Supreme Court glossed the Rules as follows:

> The first part asks about *cognitive capacity*: whether a mental defect leaves a defendant *unable* to understand what he is doing. The second part presents an ostensibly alternative basis for recognizing a defense of insanity understood as a lack of *moral capacity*: whether a mental disease or defect leaves a defendant *unable* to understand that his action is wrong.[22]

The Court did not justify the inclusion of incapacity-specific terminology. Presumably, they thought it obvious that the Rules implicated incapacities.

This incapacity-based reading is supported by a wealth of authority. English law from at least the 10th century through to the M'Naghten Rules in 1843 generally accepted that insanity ought

---

[18] I omit 'quality', as 'nature' and 'quality' are synonyms rather than independent tests: *Codere* (1916) 12 Cr App R 21.

[19] *The Queen against Daniel M'Naughton* (1843) 4 St Tr 847, 925 (emphasis added).

[20] *M'Naghten's case* (n 17) 205 (emphasis added).

[21] The editor's commentary concurs, referring to 'a mind incapable of distinguishing between right and wrong'. *M'Naghten's case* (n 17) 200.

[22] *Clark v Arizona* 548 U.S. 735; 126 S. Ct. 2709 (2006) (emphases added), quoted in RD Mackay, 'Righting the Wrong? – some observations on the second limb of the M'Naghten Rules' [2009] Crim LR 80, 80.

to be shown lenience or (later) exculpation, without explaining why, or what constituted insanity.[23] But William Hawkins' *Treatise of the Pleas of the Crown* offered the following explanation:

> The guilt of offending against any law whatsoever…can never justly be imputed to those who are *incapable* of understanding it, or of conforming themselves to it… [T]hose who are under a natural *disability* of distinguishing between good and evil, as…infants…ideots, and lunatics…are not punishable by any criminal prosecution whatsoever.[24]

This incapacity-focus followed through to the modern law. According to *Sullivan* (1984),

> "[M]ind" in the M'Naghten Rules is used in the ordinary sense of the mental faculties of reason, memory and understanding. If the effect of a disease is to impair these faculties so severely as to have either of the consequences referred to in the latter part of the rules…provided that it subsisted at the time of commission of the act [the defence is valid].[25]

As mental *faculties* are synonymous with *capacities*, *Sullivan* can be interpreted as saying that the insanity defence arises where disease causes defendants to suffer some relevant incapacities (to reason, memorise, or understand) with certain relevant effects (not knowing the nature of the act or that it was wrong). Similarly, the Law Commission implied that insanity's rationale is to account for incapacities by noting that 'it would be fundamentally unfair and unjust to hold someone criminally responsible for their conduct if, through no fault of their own, they lacked

---

[23] I include 'madness', 'idiocy', and 'lunacy' within insanity, though they were sometimes distinguished by transience. Nigel Walker, *Crime and Insanity in England: Volume 1* (Edinburgh 1968) chs 1-4. This is true at least of early Anglo-Saxon codes, as well as Bracton, Fleta, Hale, and Coke. There was a flood of legislation dealing with insanity and 'lunacy' during Victoria's reign: 35 acts of Parliament for England and Wales, plus 29 for Ireland and Scotland. But insanity was undefined throughout. Only in 1958's Divorce (Insanity and Desertion) Act is a statutory definition offered, and a contextually narrow and unsatisfactory one at that: a person 'receiving treatment for mental illness…as a resident in a hospital…'.

[24] John Curwood, *Hawkins' A Treatise of the Pleas of the Crown Vol I* (8th edn, Sweet, Maxwell and Stevens 1824) 1-2 (emphasis added).

[25] *Sullivan* [1984] AC 156, 172, citing Devlin J in *Kemp* [1957] 1 QB 399. Lord Diplock seems to have omitted a 'defect of reason' as a standalone element of the Rules, or perhaps incorporated the defect of reason into the disease of the mind. At any rate, the more traditional formulation has prevailed.

the *capacity* to obey the law."[26] In sum, the 'defect of reason' test (in *Sullivan*) and the nature and wrongness limbs (in *M'Naghten*), and sometimes both (in Hawkins' *Treatise* and the US Supreme Court) have all been given incapacity readings by the highest authorities.

This view has both comparative and reforming support. All of Germany, Belgium, Scotland, Canada, Australia, Singapore, the US Model Penal Code, and the International Criminal Court invoke capacities in their insanity tests.[27] (Among many others).[28] Similarly, though the M'Naghten Rules have seen off a regular stream of attempted reforms, none of these efforts has questioned the doctrine's incapacity-based rationale. The Atkin Committee on Insanity and Crime proposed a defence 'when the act is committed under an impulse which the prisoner was by mental disease in substance deprived of any power to resist.'[29] As with *Sullivan*'s reference to 'faculties', a 'power to resist' is synonymous with 'capacity to resist'. The later Butler Report proposed a test requiring either 'severe mental illness' or 'severe subnormality', both cashed out in terms of incapacities.[30] Finally, the Law Commission more recently proposed that

---

[26] Law Commission, *Criminal Liability: Insanity and Automatism* (Discussion Paper, 2013) [1.52] The capacity 'to obey the law' may be the underlying rationale for the more specific tests in the Rules. See ch 5.

[27] HLA Hart, *Punishment and Responsibility* (John Gardner ed, 2nd edn, Oxford 2008) 189-190: 'The German Code of 1871 spoke of *inability* or *impaired ability* to recognize the wrongness of conduct or to act in accordance with this recognition… The Belgian Loi de Défence Sociale of 1930 makes no reference to knowledge or intelligence but speaks simply of a person's *lack of ability* as a consequence of mental abnormality to control his action.'; '[I]f the person was at the time of the conduct *unable* by reason of mental disorder to appreciate the nature or wrongfulness of the conduct.' Criminal Procedure (Scotland) Act 1995 s51A (1) (inserted by the Criminal Justice and Licensing (Scotland) Act 2010 s168); '…a mental disorder that rendered the person *incapable* of appreciating the nature and quality of the act or omission or of knowing that it was wrong.' Canadian Criminal Code s16(1); '…*capacity* to understand what the person is doing, or of *capacity* to control the person's actions, or of *capacity* to know that the person ought not to do the act.' Queensland Criminal Code 1899 s27; '…*incapable* of knowing the nature of the act or that he is doing what is either wrong or contrary to law'. Singapore Penal Code s84; '…substantial *capacity* to appreciate the criminality of his conduct or to conform his conduct to the requirements of law.' American Law Institute, *Model Penal Code* (1962) §4.01. (However, the full MPC formula was only adopted in Rhode Island: *State of Rhode Island v Johnson*, 399 A 2d 469 (1969); '…a mental disease or defect that destroys that person's *capacity* to appreciate the unlawfulness or nature of his or her conduct, or *capacity* to control his or her conduct to conform to the requirements of law.' Rome Statute of the International Criminal Court (1998) Art 31 (1)(a). (All emphases added).

[28] Stanley Yeo, 'The Insanity Defence in the Criminal Law of the Commonwealth of Nations' [2008] Singapore JLStud 241.

[29] Home Office, *Atkin Committee on Insanity and Crime* (Cmd 2005, 1923).

[30] 'Severe mental illness' included among its (individually sufficient) incidents 'Lasting impairment of intellectual functions shown by failure of memory, orientation, comprehension and learning *capacity*.' Home Office, *Report of the Committee on Mentally Abnormal Offenders* [Butler Report] (Cmnd 6244, 1975) ch

> an accused should not be held criminally responsible where he or she completely lacked the *ability* to conform to the law, due to a recognised medical condition.[31]

They added that

> The defendant's lack of capacity is at the heart of our main proposal and this is its principal strength. We think the new defence will be comprehensible to the nonlawyer partly because the concept of lack of capacity is one which is employed in other contexts in the civil law and so is one with which some people are familiar. The idea of a lack of capacity leading to specific legal consequences is not new.[32]

English law past and present, various other jurisdictions, and the future envisioned by reformers all interpret insanity as an incapacity doctrine.

2.2 Infancy

The Children and Young Persons Act 1933 s50 specifies a minimum age of criminal responsibility:

> It shall be conclusively presumed that no child under the age of ten years can be guilty of any offence.[33]

Call this the infancy defence. Minimum age provisions are found in almost every other jurisdiction.[34] But they don't tell us *why* they exempt children from criminal liability. How do we know if their rationale is to account for incapacities?

18.35. (Emphasis added). Severe subnormality's definition was borrowed from the Mental Health Act 1959 as 'subnormality of intelligence…of such a nature or degree that the patient is *incapable* of living an independent life or of guarding himself against serious exploitation, or will be so *incapable* when of an age to do so.' Chs 1.13, 18.25.

[31] Law Commission, *Insanity and Automatism* (n 26) [3.1] (emphasis added).

[32] ibid [3.2] (citations omitted). They continued, 'We discuss the relevant capacities in detail [later.]' That is, they also distinguished *whether* capacities are relevant from *which*.

[33] As amended by the Children and Young Persons Act 1963 s16(1). Scottish law's provision is identical but for the age being 8 years (Criminal Procedure (Scotland) Act 1995 s41, although s41A bars prosecution for under 12s). At time of writing, Age of Criminal Responsibility Bills are in motion to raise both ages to 12.

[34] The minimum ages vary from 7 to 18. Don Cipriani, *Children's Rights and the Minimum Age of Criminal Responsibility: A Global Perspective* (Ashgate 2009) annex 2.

Consider the (now abolished) doli incapax defence. It was available for children over the minimum age of criminal responsibility but under a higher threshold (formerly 14).[35] But, unlike the minimum age rule, it was conditional: conditional on the child's *incapacity* to understand serious wrongfulness.[36] It seems highly unlikely that infancy has an entirely separate rationale from doli incapax. Hale applied precisely the same test to both defences.[37] So it is no surprise that the default modern view is that infancy is an incapacity doctrine.[38]

As with insanity, we can test this judgement against historical and comparative analogues. Ancient Hebrew and early Roman law both accounted for infancy and insanity by holding that such defendants were incapable of intentionality.[39] Later Roman law picked out more specific incapacities. Justinian's *Digest* notes that a governor in 'a state of insanity [such] that he lacks all understanding through the continuous alienation of his mental faculties' may not be punished.[40] No action arose against a lunatic for property damage as there was no 'accountable fault in him who is out of his mind', and 'the same must be said if an infant has caused

---

[35] Abolished by the Crime and Disorder Act 1998 s34, confirmed by *JTB* [2009] UKHL 20. For criticism, see Nigel Walker, 'The end of an old song?' (1999) 149 NLJ 64; Francis Bennion, 'Mens rea and defendants below the age of discretion' [2009] Crim LR 757.

[36] Stephen referred to a 'sufficient capacity to know that the act was wrong': James Fitzjames Stephen, *Stephen's Digest of the Criminal Law* (1st edn, 1877) ch III, Art 26. The Supreme Court noted that 'the common law recognised that the capacity to distinguish between right and wrong was an element of criminal responsibility…[T]he capacity to commit a criminal offence…required an ability to distinguish between right and wrong.' *JTB* (n 35) [15], [17]. *Doli*, from singular *dolus*, translates (variously) to evil, guilt, trickery, deception, mischief, etc. Lord Phillips says that 'Incapacity *of committing an offence* is as good a translation of doli incapax as any.' (*JTB* [8]). That helped to reconcile s34's wording ('incapable of *committing an offence*') with its heading (which referred to *doli incapax*). But it is a less accurate translation: the incapacity signalled by that phrase was not just to commit an offence (the *output* of the rule), but to understand wrongfulness (the *input* for the rule) (n 2).

[37] '[If] an infant be above seven years old and under twelve years…*prima facie* he is to be judged not guilty…because he is supposed not of discretion to judge between good and evil…; [but] If an infant within age be…seven years old, he cannot be guilty of felony…for *ex presumptione juris* he cannot have discretion.' Matthew Hale, *History of the Pleas of the Crown Vol I* (Nutt and Gosling 1736) 26-28.

[38] '[I]nfancy…can…be characterised as a form of deliberative incapacity defence'. AP Simester et al, *Simester and Sullivan's Criminal Law* (6th edn, Hart 2016) 692.

[39] According to Gaius, a 'lunatic cannot enter into any transaction because he does not understand what he is doing', while 'infant and those who are bordering on infancy do not differ much from insane persons, not being capable of judging for themselves'. Nor could children be guilty of theft, for that requires intentionality. Edward Poste (trs), *Institutes of Gaius* (4th edn, EA Whittuck ed, Oxford 1904) 3.106, 3.109, 3.208. Anthony Platt and Bernard Diamond, 'The Origins of the "Right and "Wrong" Test of Criminal Responsibility and Its Subsequent Development in the United States: An Historical Survey' (1966) 54 Cal LR 1227, 1228-1229.

[40] Alan Watson (ed), *The Digest of Justinian Vols 1-4* (Pennsylvania 1985) 1.18.14; see too 48.9.9. Both justify the defence on the basis that 'he is being punished enough by his very madness.' (Confinement for public protection was still permitted).

damage…[except] if the child were over seven years of age…provided the child were able to distinguish between right and wrong.'[41] That capacity of discernment, doli capaces, was evidenced by hiding after offending.[42] The law of fraud was even more explicit in exempting those of 'an age where he is…free of wrongful intent…[Though children] who are close to puberty are capable of wrongful intent…The same…must apply in the case of [the] lunatic.'[43] Roman law set a lower bound on possible liability at those incapable of speech (*infans*), and the middle ages discretion around puberty (*impubes*), later crystallizing into age boundaries of 7 and 12/14 (for girls/boys).[44] In civil law, these categories remained *de jure* until various 18th century codifications, which largely replicated the Roman categories.[45] Meanwhile, English law reinvented the same process: at first testing a child's discernment by the ability to count to twelve, or assist their parents at work, and, later, asking whether the child could distinguish good and evil.[46] All of which is evidence that infancy is an incapacity doctrine.

All these sources compared children to the insane. The comparison recurs constantly in Roman law. Gaius commented that infants 'do not differ much from insane persons'.[47] Children and 'madmen' were both subjects of *curatores*, Roman trustee-guardians. Both were barred as witnesses for wills.[48] Both were excluded from contractual obligations.[49] Both were excluded from judicial office.[50] Both were immune from liability for property damage.[51]

---

[41] ibid 9.2.5.

[42] ibid 42.7. The same test was used in medieval England: 'The 1338 Year Book says that Edward III's Judge Spigurnel decided that a child could be hanged for killing his friend, because by hiding he had shown that "he could discern between good and evil"' Walker, *Crime and Insanity* (n 23) 28.

[43] *Digest* (n 40) 44.4.4. The phrase 'capable of wrongful intent' recurs at 47.10.3.

[44] Puberty was originally set by reference the ability to procreate, later by a blanket age of 14 for boys, 12 for girls. *Institutes of Gaius* (n 39) 1.196; Thomas Collett Sandars (ed), *The Institutes of Justinian* (3rd edn, Longmans 1865) 1.22.

[45] Cipriani, *Children's Rights* (n 34) 74.

[46] Walker, *Crime and Insanity* (n 23) 28.

[47] *Institutes of Gaius* (n 39) 3.109.

[48] *Institutes of Justinian* (n 44) 2.10.6 (nor could women, slaves, or spendthrifts).

[49] *Digest* (n 40) 44.7.12-13. 'It is clear in the nature of things that a lunatic, whether he makes a stipulation or a stipulation or a promise, performs no valid act. Very near to him in position is a person who is of an age that he does not yet understand what is being done…'

[50] ibid 5.1.12: 'the permanently insane, and the *impubes* through lack of judgment are prevented by nature.'

[51] ibid 6.1.60: 'When the possessor is a child or insane and he destroys or damages something, the act is not penalized.'

Neither could give consent.[52] In each case, a single provision considers the two statuses identically.[53] Both Hale and Hawkins treated infancy and insanity under the same headings. Modern German Law uses a single term to cover both concepts.[54] By implication, their distinctive treatment is justified by their shared features. Insofar as insanity is unambiguously treated as an incapacity doctrine, so too should infancy.[55] Their rationales are mutually reinforcing.

2.3 Fitness to plead

Unfitness to plead is a procedural mirror to the insanity defence. The Criminal Procedure (Insanity) Act 1964 s4 requires a judge, on the advice of two medical practitioners, to postpone a trial if 'the accused is under a disability, that is to say, under any disability such that apart from this Act it would constitute a bar to his being tried.' The final clause clarifies that the rule accounts for a defendant's pre-existing disabilities (ie incapacities) rather than creating a legal disability. Its rationale is explicitly to account for incapacities. It's clearly an incapacity doctrine. But only peripherally. We're ultimately interested in incapacity rules as a means of accounting for less culpable defendants. Unfitness to plead need not implicate culpability. A highly culpable defendant may yet be unfit to plead if she cannot communicate.[56] That is one reason why we limited the scope of our inquiry, in chapter 1, to the substantive law. I mention it partly for completeness (unfitness to plead bars liability no less than insanity, after all),[57] but mostly because it can shed light on other incapacity doctrines.

---

[52] ibid 8.2.5: (regarding servitudes) 'against a man's wishes… [does not] mean that he openly objects, but that he does not give his consent. Thus…something can properly be said to be done against the wishes of an infant or a lunatic.'

[53] One potential exception is ibid 48.8.12: 'An infant or a madman who kills a man is not liable…the one being protected by innocence of his intent, the other excused by the misfortune of his condition.' Yet the distinction between 'innocence' and 'condition' is more naturally understood as a distinction between *aetiologies* for their shared incapacity. (See ch 4).

[54] "'*Zurechnungsfähigkeit*" (translated usually as "capacity") covers both age requirements and requirements of sufficient mental capacity. This concept…deals with that aspect of responsibility that concerns a profound ability to take responsibility for one's actions.' Kimmo Nuotio, 'On Becoming a Responsible Person' (2005) 2 Ohio St J Crim L 513, 513.

[55] One exception is James FitzJames Stephen, *A History of the Criminal Law of England* (1883) vol II 150-1.

[56] John Gardner argues that this is a false dichotomy; that 'in respect of rationale both [doctrines] are part of the same diachronic standard, which is a legal standard of basic responsibility.' 'The Mark of Responsibility' in *Offences and Defences* (Oxford 2007) 183.

[57] As pointed out by the Law Commission, *Insanity and Automatism* (n 26) [1.14].

Which incapacities render a defendant unfit to plead? They are specified only in common law. The classic formulation, from *Pritchard* (1836), asks 'whether [the defendant] is of sufficient intellect to comprehend the course of proceedings…and to comprehend the details of the evidence'.[58] The most popular modern formulation, from *M (John)* (2003), asks whether the defendant was 'capable of doing six things', which include 'understanding the charges', 'following the course of proceedings', and 'giving evidence in his own defence'.[59] The Law Commission, worried that the current test is too restrictive (among other problems), has proposed a new Criminal Procedure (Lack of Capacity) Bill, which would forbid trials where 'the defendant lacks capacity to participate effectively in the trial'.[60] The Bill specifies ten relevant abilities, including 'an ability to give instructions to a legal representative', 'an ability to make a decision about whether to plead guilty or not guilty', and the catch-all 'ability that appears to the court to be relevant'.[61] Interestingly, this proposed test shares relevant incapacities of understanding and decision-making with the insanity defence. As with infancy, this offers some mutual reinforcement of insanity's status as an incapacity doctrine (and, indeed, as to which incapacities count).[62]

2.4 Automatism

Those who act automatically do not commit the actus reus of an offence. Arguably, this is not the result of a distinct doctrine, but rather an application of the general principles of identifying actus rei.[63] But this is a miserly interpretation of automatism. Textbooks treat it as a distinct doctrine because it consists of a distinct set of rules to distinguish automatic non-offences from (on the one hand) offences and (on the other) the insanity defence. But is it an *incapacity* doctrine? There are two lines of evidence in favour.

First, automatism is often near indistinguishable from insanity. Consider two causes of dangerous driving: (1) anaphylactic shock from a bee sting; (2) an epileptic fit. Both conditions allow the sufferer to avoid liability, but via different legal routes. The first involves a non-

[58] *Pritchard* (1836) 7 C&P 303, 304.

[59] *M (John)* [2003] EWCA Crim 3452 [20].

[60] Law Commission, *Unfitness to Plead Vol 2: Draft Legislation* (2016) s1(2).

[61] ibid s3.

[62] See ch 5.

[63] Some say automatism is a denial of mens rea: Law Commission, *Insanity and Automatism* (n 26) [5.5]-[5.8]. This doesn't affect the argument.

mental cause and counts as automatism. The second is (oddly) classed as a mental cause and counts as insanity.[64] This distinction is only a matter of aetiology. It doesn't map any important functional difference.[65] So it seems highly plausible, then, that both routes reach their shared conclusion for a shared rationale: to account for incapacities.

Second, automatism is often explained in explicitly incapacity-based terms. *Hill v Baxter* (1958) equated automatism with being 'rendered unconscious or otherwise *incapacitated*'.[66] *Quick* (1973) pointed out that a 'self-induced *incapacity* will not excuse…, nor will one which could have been reasonably foreseen.'[67] By pragmatic implication, incapacities which are neither self-induced nor reasonably foreseen *may* 'excuse' by amounting to automatism. Admittedly, this is not crystal clear. *Bratty* (1963) defined the automatic person as one 'who, *though capable of action*, is not conscious of what he is doing. It means unconscious involuntary action.'[68] This seems to undermine the incapacity-based reading. But this sounds more like a point about *which* incapacity matters, not whether they do. It leaves open both that D was capable of *some* action (eg swatting bees) and that automatism involves incapacities (eg to act consciously). At any rate, more recent cases endorse broader tests than unconsciousness. As one textbook summarises,

---

[64] The famous contrast is between automatism-causing hypoglycaemia (glucose deficiency due to externally injected insulin) and insanity-causing hyperglycaemia (excess glucose due to failure to inject insulin): *Quick* [1973] QB 910; *Hennessy* [1989] 1 WLR 287. The distinction is sometimes said to be between external and internal causes. However, as the Law Commission points out, an *internal* non-mental cause (eg cramping) would also be classed as automatism. Law Commission, *Insanity and Automatism* (n 26) [5.13]-[5.15].

[65] Consider: What if you suffered a panic attack after being stung? What if the bees triggered a catatonic state due to an underlying mental illness? Would this be insanity or automatism? (The bees are from *Hill v Baxter* [1958] 1 QB 277, 282-283 and 286, citing a dictum from *Kay v Butterworth* (1945) 61 TLR 452).

[66] *Hill v Baxter* (n 65) 286 (emphasis added). Lord Goddard argued (282) that as dangerous driving required no *mens rea*, 'The justices' finding that the respondent was not *capable* of forming any intention as to the manner of driving is really immaterial.' (Emphasis added). Is this a counterpoint? No, because Lord Goddard is mistaken. It does not follow from mens rea being irrelevant that *capacity* to form mens rea is irrelevant. Some incapacities to form *mens rea* can amount to insanity, and insanity is relevant to D's liability even for strict liability offences. In context, it seems that Lord Goddard meant that if mens rea is irrelevant, *evidence as to whether D had mens rea* is irrelevant. That is correct, and the gist of the case's ratio. But formulating this in terms of D's *capacity* to have *mens rea* is misleading. I note a similar misinterpretation of Lord Goddard's regarding insanity in James Manwaring, 'Windle Revisited' [2018] Crim LR 984 (forthcoming).

[67] *Quick* (n 74) 922 (emphasis added).

[68] *Bratty v Attorney-General for Northern Ireland* [1963] AC 386, 401 (citing the Court of Appeal judgment).

> What counts is the inability. Obviously, where the defendant is altogether unconscious her reasoning capacities will be inactive. But a defendant need not be unconscious before these capacities may be suppressed or inoperative.[69]

It may be objected that a capacity being 'not responded to', 'suppressed', or 'inoperative' is distinct from being incapable. I doubt this, for reasons developed in later chapters.[70] But, even if there is a distinction, we should probably stretch the definition of incapacity doctrines to include accounting for 'inoperative' capacities. Finally, the Law Commission's provisional proposal for a statutory automatism doctrine would ask if D 'wholly lacked the capacity to control his or her conduct, and the loss of capacity was not the result of a recognised medical condition.'[71] Each source strongly implies automatism's rationale is to account for incapacities.

2.5 Diminished responsibility

Diminished responsibility (DR) is a partial defence to murder. It avoids murder's mandatory life sentence, permitting full sentencing discretion, including absolute discharge.[72] In practice, this makes it potentially as consequential as the full insanity defence. It's available where

1) D kills another, while
2) suffering an 'abnormality of mental functioning', which
3) arose from a recognised medical condition, and
4) 'substantially impaired D's ability' to understand the nature of his conduct, or form a rational judgment, or exercise self-control, and
5) provides an explanation for (ie contributed to causing) his action.[73]

Is DR an incapacity doctrine?

As with fitness to plead, DR's explicit reference to D's (substantially impaired) *abilities* implies an incapacity-accounting rationale. That view is supported, once again, by comparison with insanity. There is considerable overlap: an 'abnormality of mental functioning which…arose

[69] Simester et al, *Simester and Sullivan* (n 38) 112-113 (emphasis in original), endorsed in Law Commission, *Insanity and Automatism* (n 26) [5.2].

[70] Ch 6.

[71] Law Commission, *Insanity and Automatism* (n 26) [5.124].

[72] Homicide Act 1957 s2(3).

[73] Homicide Act 1957 s2(1) (as amended by the CJA 2009 s52).

from a recognised medical condition' essentially rephrases M'Naghten's 'defect of reason due to disease of the mind'. Further, DR accounts for more incapacities, more generously. It accounts for incapacities not only of *reason,* but also of *volition* ('exercising self-control').[74] The relevant 'abnormalities' of mental functioning are interpreted more broadly.[75] And the defendant's ability (eg) to understand the nature of their conduct need only be substantially impaired, versus M'Naghten's (supposed) requirement for complete incapacity.[76] As it accounts more generously and completely for the same incapacities identified by insanity, DR seems as clear a case as any of an incapacity doctrine.

2.6 Loss of control (provocation)

Like DR, loss of control (formerly provocation) is a partial defence to murder which reduces liability from murder to manslaughter.[77] It's available where

1) D kills another,
2) resulting from loss of self-control, due to
3) a qualifying trigger,[78] and
4) 'a person of D's sex and age, with a normal degree of tolerance and self-restraint and in the circumstances of D, might have reacted in the same or in a similar way.'[79]

Is this an incapacity doctrine? This depends on what 'loss of self-control' means. Does it mean 'loss of the *capacity* of self-control' or '*incapacity* to self-control'?

---

[74] Volitional tests within the insanity defence have been controversial in many jurisdictions, and its absence from the English insanity defence often criticised: Hart, *Punishment and Responsibility* (n 27) 189.

[75] It includes 'not only the perception of physical acts and matters and the ability to form a rational judgment whether an act is right or wrong, but also the ability to exercise will-power to control physical acts in accordance with that rational judgment'. *Byrne* [1960] 2 Q.B. 396, 403. See RD Mackay, 'The abnormality of mind factor in diminished responsibility' [1999] Crim LR 117, and RD Mackay and Barry Mitchell, 'The new diminished responsibility plea in operation: some initial findings' [2017] Crim LR 18.

[76] For a while, following *Brown* [2011] EWCA Crim 2796, 'substantial' was interpreted as anything more than merely trivial. However, the Supreme Court has since clarified that substantial means 'weighty'. *Golds* [2016] UKSC 61. It's commonplace that *M'Naghten* requires total incapacity (Law Commission, *Insanity and Automatism* (n 26) [3.3]), but I'll question what this means in ch 6.

[77] CJA 2009 ss54-56.

[78] The qualifying triggers are specified in s55 as 'fear of serious violence from V against D or another identified person', or something which 'constituted circumstances of an extremely grave character, and…caused D to have a justifiable sense of being seriously wronged', or both.

[79] CJA 2009 s54(1).

Both DR and loss of control are governed by the Coroners and Justice Act 2009.[80] DR (s52) was concerned (inter alia) with a substantially impaired *ability* to exercise self-control. Loss of control (s54) shares the language (and presumably concept) of self-control. On one reading, then, perhaps the two provisions share an underlying concept that is simply more fully explicated in DR: an *ability* to (exercise) self-control. Then loss of control sounds like an incapacity doctrine. Call this the *incapacity reading*. On an alternative reading, perhaps the wording differs between the two (nearly adjacent) sections precisely to distinguish the underlying concepts: DR's *ability* to exercise self-control is not the same as a *loss* of self-control. Call this the *non-capacity reading*. Prima facie, the non-capacity reading seems more plausible. But we could explain the different formulations without invoking a different underlying concept. Loss of control must be total. Incapacity-based formulations ('a complete loss of capacity to self-control') imply an *ongoing* or *permanent* incapacity. But that is not required: transient loss of control suffices. A capacity-free 'loss of control' formulation might have been used to clearly include transient (if total) losses of control, even if the underlying concept was of an incapacity to self-control. (DR's 'substantial impairment' of ability poses no such problems of clarity). The different formulations are explicable without relying on the non-capacity reading. Ultimately, however, the statutory formulation doesn't tell us which reading is correct.[81]

The courts haven't clarified much. The leading judgment said only that loss of control's meaning was 'self-explanatory'.[82] By contrast, the Law Commission claimed that '[t]here is no satisfactory definition of loss of self-control' and that 'the requirement of loss of self-control …[lacks] sharpness or a clear foundation in psychology'.[83] There is only a single judicial gloss from the old law, from *Oneby* (1727). It defined loss of control as 'such a passion as for the time deprives [D] of his reasoning faculties'.[84] This (weakly) suggests the incapacity reading.

---

[80] Both spurred by the Law Commission, *Partial Defences to Murder* (Final Report, 2004).

[81] Another reason to worry about the statutory formulation is that the drafting is very unclear in other regards. Much criticised is how s55(6)(c) (partially) excludes sexual infidelity from consideration: see *Clinton* [2012] EWCA Crim 2 [11]-[28], and the ambiguity as to whether intoxication counts as a 'relevant circumstance', contra the old provocation defence (the idea was finally kiboshed in *Asmelash* [2013] 1 Cr App R 33).

[82] *Clinton* (n 81) [10].

[83] Law Commission, *Partial Defences to Murder* (n 80) [3.26] and [3.28] (the discussion is about the old provocation defence, but the point is unchanged).

[84] *Oneby* (1727) 92 ER 465, via ibid [3.26].

Aside from limited authorities, there are two potential objections to the incapacity reading. First, the Law Commission offered a conceptual objection:

> The term loss of self-control is itself ambiguous because it could denote either a *failure* to exercise self-control or an *inability* to exercise self-control. To ask whether a person *could* have exercised self-control is to pose an impossible moral question. It is not a question which a psychiatrist could address as a matter of medical science.[85]

Oddly, however, the Law Commission later in the same report proposed the test for DR mentioned above: that D's *capacity* to exercise self-control was 'substantially impaired'. It's not clear why a test for the substantial impairment of the capacity to self-control is acceptable while a test for the incapacity to self-control would pose an impossible moral question.[86] This sounds like unwarranted selective scepticism.[87]

The second objection to the incapacity reading is that s54(1)(c) explicitly requires that defendants be judged as if they had a 'normal degree of tolerance and self-restraint', while s54(3) adds that

> the reference to "the circumstances of D" is a reference to all of D's circumstances other than those whose only relevance to D's conduct is that they bear on D's *general capacity* for tolerance or self-restraint. [Emphasis added.]

The motivation for this section is clear: some (eg) personality disorders result in violent dispositions and lack of self-restraint. These disorders will often result in D 'losing control'. But Parliament thought it a mistake to extend greater lenience to such defendants in virtue of their volatile dispositions.[88] The objection is that it's hard to see the distinction between a 'general capacity…for self-*restraint'* and a 'capacity to self-*control'*. They sound essentially identical. If they are, D's capacity to self-control is ruled out as irrelevant to the defence. If so, loss of control cannot refer to 'an incapacity to self-control'. And if so, loss of control is not an incapacity doctrine. But there *is* room to distinguish self-restraint from self-control. Why

---

[85] Law Commission, *Partial Defences to Murder* (n 80) [3.28] (emphasis added).

[86] ibid [5.97]. Admittedly, the DR proposal was tentative ([5.92]-[5.93]) and only required that the defendant 'acted in response to' a provocation, not that they lost control. ([3.168]).

[87] While the concept of the capacity to self-control is certainly a difficult philosophical matter (see chs 3, 5, and 6), so too are ubiquitous legal test of intention, causation, etc.

[88] See the Explanatory notes and *Rejmanski* (n 9).

else would the statutory formulation rule out the capacity for 'tolerance' if self-restraint did all the work?[89] Similarly, the addition of the qualifier 'general' can distinguish between (excluded) long-term incapacities (like personality disorders) from (permitted) momentary incapacities (due to provocation) as a proper basis for loss of control. And, as noted above, we're only asking *whether* incapacities are relevant, not *which*. Once again, the statutory formulation is inconclusive.

Given the limited evidence, any conclusions must be speculative. The only clear authority, *Oneby*, favours the incapacity reading. It seems (just about) more likely that loss of control is an incapacity doctrine than not.

3 Relativisations

Recall that incapacity relativisations are standards relativised (perhaps implicitly) to defendants' incapacities, and that in identifying them we must avoid merely incidentally relevant incapacities.

3.1 Duress: reasonable firmness

Duress is a complete defence to almost all criminal offences.[90] Though ancient in origin, successful pleas were rare until the 1970s.[91] The exact test wasn't clear until *Graham* (1982).[92] Including subsequent refinements, it is:

1) The defendant was 'impelled to act as he did because, as a result of what he *reasonably believed* [a threatener/circumstances[93]] had said or done, he had good

[89] Admittedly, the M'Naghten Rules refer to D's understanding of the 'nature and quality' of his conduct, but these are not distinct tests.

[90] Many critics, including Stephen, the Law Commission, and various judges argue that duress would make more sense as a mitigating factor or partial defence rather than a complete defence. Because of this, the law has been interpreted very strictly. It is not a defence to (attempted) murder: *Howe* [1987] AC 417. (And possibly treason).

[91] *Z/Hasan* [2005] 2 AC 467 [22], noting that neither Stephen (in 1883) nor Hart (in 1960) knew of many cases.

[92] *Graham* [1982] 1 WLR 294. The test was endorsed by the Lords in *Howe* (n 100).

[93] Duress by circumstances was acknowledged in *Martin* (1989) 88 Cr App R 343.

cause to fear that if he did not so act [they/it] would kill [or seriously harm[94]] him [or another]'

2) [A] sober person of *reasonable firmness*, sharing the characteristics of the defendant…[might] have responded to whatever he *reasonably believed* [the threatener/circumstances] said or did by taking part in the [offence][95]

The two 'reasonable belief' standards could each be relativised.[96] But we'll focus on the core, distinctive standard of duress: reasonable firmness in resisting the threat/circumstance. Is this relativised to account for D's incapacities?[97]

We first need to know whether 'the characteristics of the defendant' include her incapacities. *Graham* concerned the following jury direction:

> Taking into account all the circumstances of the case, including the age, sex, sexual propensities and other characteristics personal to the defendant, including his state of mind and the amount of drink or drugs he had taken, was it reasonable for the defendant to behave in the way he did[?]… The test of reasonableness in this context is: would the defendant's behaviour in all the particular circumstances…reflect the degree of self-control and firmness of purpose (which) everyone is entitled to expect that his fellow citizens would exercise in society as it is today[?][98]

The prosecution thought the test too lenient: they argued the standard should not account for D's drink and drug use. They succeeded.[99] But the Court of Appeal *only* objected to the trial judge's mention of self-induced drink and drugs. By omission, this perhaps offers an implicit

---

[94] ibid 345.

[95] Technically, the second limb was phrased negatively, such that the defence is *not* available if the reasonable person would *not* have responded as D did. *Graham* (n 92), 298. To generate a positive formulation, I've interpreted this to mean that the defence *is* available if the reasonable person 'might' have so responded.

[96] There are yet further standards. The defence is denied to defendants who voluntarily associated with their threateners (eg a gang), if it was *reasonably foreseeable* they'd suffer duress-inducing threats as a result. Softening that rule, D is (possibly, it was *obiter*) permitted a *reasonable excuse* for such associations (eg if the threatener is a partner or family member). (*Z*/*Hasan* (n 91) [39], [77]-[78]). Given their recent origins, there is no guidance as to whether these standards may be relativised.

[97] Is duress an incapacity *doctrine*? Possibly: a common gloss of the test is whether the defendant lost their '*capacity* to act independently' of the threatener or had 'the freedom of will to act independently' (*Emery* (1993) 14 Cr App R (S) 394, 395, 396). But we'll not pursue the thought.

[98] Quoted in *Graham* (n 92) 298.

[99] ibid 300.

endorsement of the relevance of 'age, sex, sexual propensities…other characteristics…[and] state of mind…in all the particular circumstances.'[100] But this is very weak evidence.

Several further cases tackled the issue head-on. The route to appeal was always the same: D claimed duress and sought to admit specialist psychiatric (psychological, medical) evidence to demonstrate some incapacity. Such specialist evidence is admissible only if relevant to D's liability.[101] The courts had to decide whether various incapacities were relevant to reasonable firmness. At first, they supported *Graham's* (implicitly) liberal relativisation. In *Emery* (1993), the court accepted evidence of battered women's syndrome, a subset of post-traumatic stress disorder (PTSD), as relevant to reasonable firmness.[102] But the approach soon hardened. In *Hegarty* (1994), D tried to admit evidence of mental instability.[103] In *Horne* (1994), D was unusually pliable.[104] And in *Hurst* [1995], as with *Emery*, the evidence regarded heightened compliance due to abuse.[105] None succeeded. In each case, the judges argued that reasonable firmness couldn't account for evidence of a defendant's attributes which connoted a *lack* of firmness. The test couldn't account for 'emotional instability' or a 'grossly elevated neurotic state' (*Hegarty*); nor 'mental characteristics such as inherent weakness, vulnerability and susceptability [sic] to threats' (*Horne*); nor any 'characteristics of a personality which lacks reasonable firmness' (*Hurst*). The problem, however, is that *young age* may be a cause of 'emotional instability', 'vulnerability', or a 'characteristic of a personality which lacks reasonable firmness'. But all these cases affirmed that D's age (and sex, and serious physical disabilities)

[100] The Lords of Lords only reiterated that reasonable firmness was to be judged relative to 'the characteristics and in the situation of the defendant', adding that 'No doubt there are subjective elements…' *Howe* (n 90) 426.

[101] The evidence also must be outside the knowledge of laypeople (the jury): *Turner* (1974) 60 Cr App R 80.

[102] E was convicted of child cruelty for failing to stop her partner, D, beating and ultimately killing their infant child. E claimed duress by D, who appealed against the admissibility of evidence of PTSD (which affected his case). *Emery* (n 97), 398-399. The Court did accept that the evidence went beyond its proper remit, however, as it strayed into matters within the knowledge of the jury: *Turner* (n 101).

[103] *Hegarty* [1994] Crim LR 353 (regarding a robbery). While the primary issue was apparently whether D's mental condition affected his response to threats (*reasonable firmness*), it seems more plausible that his condition affected his (reasonable?) *belief* in the threats, which the prosecution argued to be non-existent. While the case report is somewhat ambiguous on the point, I will assume the primary issue was indeed of reasonable firmness.

[104] *Horne* [1994] Crim LR 584 (regarding a dole fraud conspiracy).

[105] *Hurst* [1995] 1 Cr App R 82 (regarding cocaine smuggling).

remained relevant to reasonable firmness.[106] It remained unclear which attributes D's 'characteristics' included.

The Court of Appeal tried to clarify matters in *Bowen* (1997), where the evidence was of D's extremely low intelligence.[107] They endorsed the gist of the restrictive cases.[108] But they accepted the relevance of certain attributes:

> The defendant may be in a category of persons whom the jury may think less able to resist pressure than people not within that category. Obvious examples are age, where a young person may well not be so robust as a mature one; possibly sex, though many women would doubtless consider they had as much moral courage to resist pressure as men; pregnancy, where there is added fear for the unborn child; serious physical disability, which may inhibit self protection; [and] recognised mental illness or psychiatric condition, such as post-traumatic stress disorder leading to learnt helplessness.

Evidence of mental illness, the Court held, is admissible

> provided persons generally suffering from such [a] condition may be more susceptible to pressure and threats… [But it] is not admissible simply to show that in the doctor's opinion a defendant, who is not suffering from such illness or condition, is especially timid, suggestible or vulnerable to pressure and threats

And as for D, the Court did

> not see how low I.Q., short of mental impairment or mental defectiveness, can be said to be a characteristic that makes those who have it less courageous and less able to withstand threats and pressure.

Bowen's key distinction is to accept the relevance of attributes which fall within *categories* of people less able to resist pressure, but to reject the relevance of attributes that tell of the 'mere fact' of pliability, vulnerability, timidity, or susceptibility. The relevant categories must *correlate*

---

[106] *Horne* (n 104) 585.

[107] *Bowen* [1997] 1 WLR 372, 379-380 (regarding credit fraud).

[108] ibid 379-380: 'The mere fact that the defendant is more pliable, vulnerable, timid or susceptible to threats than a normal person is not a characteristic with which it is legitimate to invest the reasonable/ordinary person'.

with, but cannot be *defined by*, threat-susceptibility.[109] But does this mean reasonable firmness is relativised? And if so, is it relativised (perhaps implicitly) to incapacities?

Let's take the questions backwards. *If* the standard is relativised, it's an incapacity relativisation. *Bowen* asks explicitly whether defendants fall within 'a category of persons…*less able* to resist pressure', those '*less able* to withstand threats.' It distinguishes between mental-illness-induced 'susceptibility' (relevant) and non-mental-illness-induced 'timidity' (irrelevant), a distinction which plausibly tracks the extent to which those (seemingly identical) attributes involve incapacities.[110]

But *is* the standard relativised? *Bowen* explains the relevance of pregnancy in terms of fear for the unborn child, and the relevance of physical disability in terms of inhibited self-protection.[111] These are merely incidentally relevant incapacities. Pregnancy and physical disability make the gravity of threats more severe. It would be reasonable for the pregnant or physically disabled person to succumb to threats more easily than non-pregnant or able-bodied people, without altering the standard of reasonableness.[112] However, mental illness can't be merely incidentally relevant. A threat to someone suffering PTSD and learned helplessness is not graver for that reason; it would not result in greater harm, as it might for children, the pregnant, or the physically disabled. Rather, we are entitled to ask for less (in terms of resisting threats) from those suffering PTSD. Including mental illness as a determinant of reasonable firmness implies relativisation, at least to some attributes.[113] That message is reinforced by the final relevant case. In *Antar* (2004),[114] D tried to admit evidence of his 51 IQ and very high suggestibility. Despite *Bowen* rejecting IQ (short of 'mental impairment') as irrelevant, the court accepted that D's low IQ and other diagnoses amounted to a 'lack of mental capacity', relevant

---

[109] The category cannot be defined by threat-susceptibility because it's trivial to construct an arbitrary category of 'persons incapable of resisting' composed of the n% most timid individuals. It is tempting to say that if D is *incapable* of resisting he may benefit from the defence, whereas if he is *capable* but merely *disposed* not to resist he may not. (*Emery* (n 97) 395-396: duress involves D lacking the '*capacity* to act independently' of the threatener or 'the freedom of will to act independently'.) But if D is sufficiently pliable, vulnerable, etc., it is plausibly correct to say that D *is* incapable of resisting, not merely disposed not to resist. (Eg per John Gardner and Timothy Macklem, 'Compassion without Respect? Nine Fallacies in R. v. Smith' [2001] Crim LR 623, discussed ch 3). That would contradict *Bowen* (n 107).

[110] Plus, the 'robustness' of age may include corollary incapacities.

[111] The same explanation for pregnancy's relevance is offered in *C* [2013] EWCA Crim 1472 [33].

[112] Then perhaps the (lack of) robustness of youth should be interpreted as merely incidental too.

[113] *Bowen* (n 107) itself doesn't distinguish relativisations from incidentally relevant incapacities.

[114] *Antar* [2004] EWCA Crim 2708 (D and two others robbed three youths of their mobile phones; D claimed duress from the other two).

to reasonable firmness.[115] Once again, we have explicit accounting for incapacities.[116] Most interestingly, D's high score on the Gudjonsson Suggestibility Scale was (emphatically) held to be admissible.[117] Suggestibility does not make threats worse. It just makes it more likely that D would succumb.[118] This again implies relativisation of the standard, not incidentally relevant incapacities.

These conclusions are tentative. Further cases may restrict the scope of the relevant categories or harden their boundaries. Indeed, because duress is a complete defence where many would prefer it to be partial, or a mitigating factor in sentencing, the courts have openly restricted and restrained its scope.[119] But the current authority strongly suggests that reasonable firmness in duress is an incapacity relativisation, and a liberal one at that.

3.2 Negligence: reasonable care

The mental element for many offences consists of failing to live up to the standard of a reasonable person.[120] Formulations differ between offences. It is careless driving to fail to demonstrate the care of 'a competent and careful driver'.[121] It is harassment to cause others to fear violence on at least two occasions in circumstances where 'a reasonable person…would think the course of conduct…would cause the other…to fear [violence].'[122] Minor differences aside, however, these offences are grouped together as *negligence* offences. It is negligent to fall

[115] D's attributes amounted to a 'moderate' or 'mild' learning disability, according to the expert psychologist.

[116] The court referred to mental 'capacity' ('that is, mental handicap') in preference to the stronger notion of 'mental illness': *Antar* (n 114) [33], borrowing and elucidating the term 'mental defective' from *Masih* [1986] Crim LR 395. This helpfully extends *Bowen's* 'category' approach to include extreme ends of scalar properties like intelligence.

[117] *Antar* (n 114) [41], [46]: '*importantly*…he had a level of suggestibility sufficiently higher than that of the general population' (emphasis added).

[118] The use of the Suggestibility Scale is curious, as it seems to measure precisely the sort of 'mere fact' of pliability that *Bowen* rejected as irrelevant. If its use remains admissible, this implies a strong liberalisation of what counts as a relevant incapacity.

[119] *Z*/*Hasan* (n 91).

[120] Gross negligence manslaughter and public nuisance are the only common law offences; the rest are statutory.

[121] Road Traffic Act 1988 ss3, 3ZA.

[122] Protection from Harassment Act 1977 s4(2).

below the standard of a reasonable person.[123] Is this standard relativised to account for D's incapacities?

Usually, the standard is explicitly relativised to D's *knowledge*. Harassment's reasonable person possesses 'the same information' as D.[124] D might have information that makes otherwise unreasonable beliefs reasonable (it is prima facie harassing to bombard V with horrific stories, but potentially reasonable if informed that V enjoyed them).[125] If D lacks relevant information due to an incapacity, eg a delusional failure to appreciate V's objections, is this an implicit incapacity relativisation? The courts have rejected the premise: delusional beliefs cannot be reasonable.[126] What about relativisations to other attributes? The cases are divided. The reasonable driver is *not* relativised to D's *skill* (as per tort).[127] The standard of negligence in harassment is *not* relativised to D's *schizophrenia*.[128] But the reasonable armed serviceman *is* relativised to D's *training* and *experience*.[129] The standard of negligence in animal cruelty *is* (possibly) relativised to D's *age* and *household seniority*.[130]

---

[123] The standard also differs qualitatively. While careless driving is negligent, it is *grossly* negligent—dangerous driving—if 'it would be obvious to a competent and careful driver that driving in that way would be dangerous'. Road Traffic Act 1988 ss3, 3ZA.

[124] Protection from Harassment Act 1977 s4(2).

[125] D could also have information that makes otherwise reasonable conduct unreasonable: it is reasonable to invite V for lunch, but potentially harassing if informed that V does not want to be contacted. Traffic offences account for capacities *only* to hinder D's case. The reasonable driver is relativised to 'the circumstances [D]…could be expected to be aware' of and 'any circumstances shown to have been within [D's] knowledge', but *not* to D's ignorance: Road Traffic Act 1988 s3ZA(3). Given this asymmetry, it won't be an incapacity relativisation.

[126] *B(MA)* (n 12).

[127] *Bannister* [2009] EWCA Crim 1571, following the lead from tort in *Nettleship v Weston* [1971] 3 WLR 370.

[128] *C (Sean Peter)* [2001] EWCA Crim 1251 [17]-[19].

[129] *Price* [2014] EWCA Crim 229 [20].

[130] *RSPCA v C* [2006] EWHC 1069 (Admin) [15]-[16]. C's case was decided under the Protection of Animals Act 1911 s1(a), which prohibited 'cruelly' ill-treating animals and 'unreasonably' permitting suffering. This was generally held to require advertence to the suffering. C was held not to have unreasonably omitted to assist her injured cat, given that she, a 15-year-old girl, had deferred to her father's decision not to seek treatment for their jointly-owned cat. The standard of reasonableness was relativised to her age and household seniority. But the 1911 act was subsequently superseded by the Animal Welfare Act 2006. Section 9(1) omits any reference to cruelty. It has been interpreted not to require advertence: *R (Gray) v The Crown Court Aylesbury* [2013] EWHC 500 (Admin). Toulson LJ noted the difference between the acts and glossed the latter as 'setting a purely objective standard of care' (at [31]). This is correct insofar as advertence is no longer required. But it shouldn't imply no relativisation at all. The main defendant in *Gray* was an adult professional horse trader: his case did not raise the issues of relativisation raised in *C*. While Gray's 16-year-old son was a co-defendant, he did not appeal this

This is a ragtag band of attributes and outcomes. Any general explanations will be sketchy. But here are two possibilities. Possibility one: the different outcomes may be explained by differences between the *offences*. The Protection from Harassment Act has a 'protective and preventative' policy.[131] It's easy to avoid harassing others. So it makes sense not to relativise its reasonable person, if doing so would weaken those protections. By contrast, the Armed Forces Act makes the negligent performance of *any* duty an offence.[132] It's difficult for armed servicemen to avoid occasional slips in their working environment. So it makes sense to relativise its reasonable person to account for these difficulties. The differences between harassment and armed service duties explain the differences in relativisation. Possibility two: the different outcomes may be explained by differences between the (would be) relativised *attributes*. *None* of the reasonable driver, harasser, or armed serviceman are relativised to the defendants' level of skill.[133] So perhaps the permissible relativisations—to training, experience, seniority, and age—would be accepted for all or most offences. (It just so happens that drivers cannot rely on low training/experience/age given that they must be of age and pass a test, and these attributes do not bear on harassment.) In other words, relativisations to certain attributes are invariant across offences, an outcome obscured by supervening considerations (eg that drivers are imputed to be sufficiently trained by completing a test). I suspect that both possibilities explain the differences between the offences to some degree.

What follows? One feature of the permissible relativisations—to training, experience, seniority, and age—is that these attributes are not within the defendant's control (at the time of the offence). By contrast, an impermissible relativisation—to demonstrate skill—is (arguably) within their control. So perhaps relativisation to uncontrollable factors is permitted. If so, the standard of reasonable care in negligence must be an incapacity relativisation: incapacities are paradigmatic examples of uncontrollable attributes. The evidence is shaky at best. But this view does have supporters. *Simester and Sullivan* endorses relativisation to age, gender, physical characteristics, and intelligence, arguing that '[t]he reasonable man test should be subjective to the extent that the defendant's shortcomings do not disclose fault'. Which shortcomings do not disclose fault? 'Academic argument favours taking account of personal

point. Further, unlike C, Gray Junior acted in a professional capacity in a case involving systemic and severe animal abuse.

[131] *C (Sean Peter)* (n 128) [17].

[132] Armed Forces Act 2006 s15(2).

[133] *Price* (n 129) [20].

incapacities.'[134] In this, they follow the approach of HLA Hart and the Canadian Supreme Court.[135]

This argument faces difficulties. The paradigmatic case of incapacity relativisation is perhaps relativisation to mental illness, of which schizophrenia is the most visible case. But the courts have rejected schizophrenia-based relativisations. Further, the persuasive authority in favour of the incapacity relativisation interpretation of the case from one leading textbook and the Canadian Supreme Court is matched by opposing persuasive authority from another leading textbook and the Australian High Court.[136] But there is hope. The authorities rejecting incapacity relativisation either qualify the approach or do so on policy grounds at odds with a culpability principle.

The Australian approach is qualified. Kirby J accepted that 'it would not be rational to impute blame to a person who is physically or mentally incapable of achieving the standard of care expected by the criminal law.' But the case concerned only *aggravated* negligence, which, Kirby J argued, preserves a correlation between blameworthiness and liability: in 'the overwhelming majority of cases, a person who causes death by aggravated criminal negligence will be regarded as extremely blameworthy.'[137] In other words, Kirby J (reasonably) did not object to the bare legal possibility that incapable, non-culpable defendants will be convicted. He objects only if they are in fact convicted.[138] But if the correlation is not high his conclusion is false: non-culpable, incapable defendants will be convicted. Then presumably Kirby J would favour relativising the standard to incapacities to avoid that outcome.

The English refusal to relativise to schizophrenia in harassment requires the rejection of a culpability principle. In *C (Sean Peter)* (2001), Hughes J emphasised the 'protective and preventative' focus of the harassment act. He worried that relativising the act's reasonableness standard to schizophrenia

---

[134] Simester et al, *Simester and Sullivan* (n 38) 162-166 (emphasis added).

[135] HLA Hart, 'Negligence, Mens Rea and Criminal Responsibility' in his *Punishment and Responsibility* (n 27); *Creighton* (1993) 105 DLR (4th) 632. While the majority in *Creighton* disagreed with Lamer CJ's dissent as to the extent of permissible relativisations, they accepted some incapacity relativisations.

[136] David Ormerod and Karl Laird, *Smith and Hogan's Criminal Law* (14th edn, Oxford 2015) 6.1.2.2; *Lavender* [2005] HCA 37 [128].

[137] *Lavender* (n 136) [128].

[138] See ch 1.

> would be to remove from its protection a very large number of victims… [and] exclude not only suitable punishment for the perpetrator, but also damages, and, more especially, an injunction or restraining order for the protection of the victim.[139]

He added that relativisation is appropriate in provocation and duress as that involves

> the question [of] when a particular defendant is to be excused from responsibility, notwithstanding that the constituents of the offence are made out… [whereas in negligence] the question is whether the constituents of the offence are made out.[140]

These policy concerns are sensible. Relativising negligence would leave the courts without protective measures for victims of mentally disordered harassers, as those powers are conditional on D offending, which D would not if negligence were relativised. But mark what this means: the law requires the advertent conviction of non-culpable defendants. It is unjust to convict a 10-year-old of harassment for conduct they could not appreciate would be harassing, even if doing so conferred a power to protect the 10-year-old's victim. Likewise, for schizophrenic defendants. Perhaps this is worth the trade-off. But not if there is a better option.[141] There is. We could provide protection that is not conditional on convicting non-culpable defendants. We could still use the criminal law. We could use a special verdict.[142] Given this incompatibility with a culpability principle, we should be very careful to endorse harassment's strict approach, let alone extend it to other negligence standards.

Ultimately, only some negligence standards seem to be incapacity relativisations. The authorities are split. I have suggested that, if we take a culpability principle seriously, we should favour the incapacity relativisation interpretation. But further judgments are required to see if this interpretation prevails.

3.3 Sex offences: reasonable beliefs

---

[139] *C (Sean Peter)* (n 128) [19]

[140] ibid [24]

[141] While some schizophrenic defendants could avoid conviction by pleading insanity, this option is extremely restricted.

[142] See ch 1 §3. The Canadian Charter of Rights and Freedoms s7 ensures that 'Everyone has the right to life, liberty and security of the person and the right not to be deprived thereof except in accordance with the principles of fundamental justice.' The Canadian Supreme Court has held that one principle of fundamental justice is 'to ensure that the morally innocent not be punished.' *Creighton* (n 135) 17d-f, citing *Re B.C. Motor Vehicle Act* [1985] 2 SCR 486, 496 and *Vaillancourt* [1987] 2 SCR 636.

The Sexual Offences Act 2003 makes it an offence for A intentionally to engage in various sexual acts with B if B does not consent and A does not '*reasonably believe'* B consents.[143] 'Whether a belief is reasonable is to be determined having regard to all the circumstances, including any steps A has taken to ascertain whether B consents.'[144] It's also an offence to engage in various sexual acts with a child between 13 and 15 years old, regardless of consent, if one does not '*reasonably believe*' they are at least 16.[145] Are these standards of reasonable belief relativised to D's incapacities?[146]

Only two post-2003 Act cases have clarified the nature of the reasonableness requirement, and in both the clarification is obiter.[147] First, in *M(M)* (2011), D claimed that his bipolar affective disorder was relevant to the reasonableness of his belief in V's consent.[148] His appeal was dismissed, as D chose not to rely on the psychiatric evidence, which at any rate contradicted his claim that V consented.[149] But the court did briefly note, regarding the meaning of a reasonable belief in 'all the circumstances', that

> it is arguable that the circumstances may include a mental illness which materially affected the defendant's ability to interpret the complainant's lack of consent.

But it was 'not the time to engage in that argument'.[150]

---

[143] Sections 1(1) (rape), 2(1) (assault by penetration), 3(1) (sexual assault), 4(1) (causing a person to engage in sexual activity). The definition of 'sexual' in s78 further invokes a 'reasonable person's' considerations.

[144] Sections 1(2), 2(2), 3(2), 4(2). The Act further specifies circumstances giving rise to legal presumptions that D did not reasonably believe in consent at ss75-76.

[145] Sections 9-12. The maximum sentence is lower for defendants who are themselves under 18 (s13). There is no reasonable belief condition for corresponding intentional acts with a child under 13 years old (ss5-12).

[146] Recall that duress requires a reasonable belief in the existence of a threat. Our conclusions about reasonable beliefs here probably transfer to that context.

[147] In *Grewal* [2010] EWCA 2448 it was argued that D's belief should be relativised to intoxication. Predictably, this was rejected, in line with the law's general approach to intoxication ([29]-[31]) in *Majewski* [1977] AC 443 and *Heard* [2008] QB 43. The pre-2003 Act approach, from *DPP v Morgan* (n 12), was that a genuine belief in consent sufficed.

[148] *M(M)* [2011] EWCA Crim 1291. D's girlfriend had recently split up with him. Despite this, he entered her parents' house, where she lived, and had sex with her. His main claim was that V consented.

[149] The expert psychiatrist thought that D was aware that V did not consent, but thought her refusal, on the grounds that her parents were in the house, was unjustified. Obviously, these findings did not favour D's case. The psychiatrist also deemed D to be legally insane, which D chose not to rely on.

[150] *M(M)* (n 148) [54].

Second, in *B(MA)* (2013), D claimed that the standard of reasonable belief in consent should be relativised to account for his paranoid schizophrenia.[151] The court held that a delusional belief in V's consent could not have been reasonable, for that 'would be, by definition, irrational and thus unreasonable, not reasonable'.[152] It placed significant weight on the fact that the Sexual Offences Act deliberately rejected the earlier approach of accepting D's genuine belief in consent, approvingly citing David Ormerod's argument that [153]

> the ease with which the defendant can ascertain the consent of his partner, coupled with the catastrophic consequences for the victim if the defendant acts without consent, militate strongly against the purely subjective approach.[154]

On the flip side, these factors militate in favour of a 'purely objective', that is, non-relativised approach. Thus, it would have been unsurprising for the court to reject all relativisations. But this they did not do. They continued:

> It does not follow that there will not be cases in which the personality or abilities of the defendant may be relevant to whether his positive belief in consent was reasonable. It may be that cases could arise in which the reasonableness of such belief depends on the reading by the defendant of subtle social signals, and in which his impaired ability to do so is relevant to the reasonableness of his belief… Whether (for example) a particular defendant of less than ordinary intelligence or with demonstrated inability to recognise behavioural cues might be such a case, or whether his belief ought properly to be characterised as unreasonable, must await a decision on specific facts.[155]

By noting an 'arguable' case for relativising to other mental illnesses and sketching out a detailed candidate involving incapacities to read behavioural cues, the court strongly implies that courts *will* relativise in appropriate cases. If they do, it will surely be a relativisation to

---

[151] *B(MA)* (n 12). D had non-consensual sex with his partner, alongside a series of bizarre antics, including forcing her to drink canned peas with crushed apple leaves. Common to other schizophrenics, D believed he had solutions to grand global problems and fantastical healing powers, including powers of sexual healing. As in *M(M)*, D's main claim was that V consented

[152] *B(MA)* (n 12) [35]. Again per M(M), D's delusions didn't mean he believed V consented, only that she *ought* to have consented, given his healing powers. Again, this wouldn't suffice to excuse him even if true: [20], [34].

[153] ibid [36]-[37].

[154] ibid [37], citing David Ormerod, *Smith and Hogan's Criminal Law* (13th edn, Oxford 2011) 744.

[155] *B(MA)* (n 12) [41].

incapacities. Note the examples given: the '*abilities*' of the defendant; an '*impaired ability*' to read subtle social signals; an '*inability*' to read behavioural cues, perhaps for reasons of low intelligence.

This conclusion is necessarily tentative. We can only clearly draw an upper bound on possible relativisations: a reasonable belief in V's consent is *not* relativised to disorders resulting in 'irrational' beliefs. We don't have a lower bound. Both cases leave open whether any relativisation is permissible. But each strongly supports some limited incapacity relativisation of the standard of reasonable belief.

4 Counterfactual relativisations

Recall that counterfactual incapacity relativisations are rules whose rationale for the *absence* of a standard (which could plausibly have been imposed) is to account for incapacities.

4.1 Recklessness: advertence to risk

Recklessness is the mental element in many offences, including assault, battery and criminal damage. The precise test differs between offences.[156] But the dominant interpretation is that D:

1) adverted to a risk of harm, and
2) unreasonably took that risk

The standard of reasonable risk-raking is not relativised. The first limb does not impose a standard (eg that D should have adverted to the risk). Is this a counterfactual incapacity relativisation? The answer is clearly yes.

For twenty years, between the landmark decisions in *Caldwell* (1982) and *G* (2003), the dominant interpretation of the first limb was that it set a standard of *reasonable* inadvertence to

[156] The famous decisions in *Caldwell* and *G* (both n 12) both concerned property damage. But their interpretations of recklessness have usually been held to apply more broadly, eg to assault (*Savage, Parmenter* [1992] 1 AC 699), pre-2003 Act sex offences (*B (a minor) v DPP* [2000] 2 AC 428), niche statutory offences (*Foster v CPS* [2013] EWHC 3885 (Admin), concerning an offence of recklessly destroying badger setts) and more. Driving offences are the exception: *G* [28] explicitly refused to overturn the *Caldwell* approach taken in *Lawrence (Stephen)* [1982] AC 510. At any rate, the Road Traffic Act 1988 refers to 'dangerous' driving, defined in terms of a 'careful and competent' driver rather than recklessness.

a risk.[157] A long series of defendants in a long series of cases challenged that interpretation in the higher courts, all unsuccessfully until *G*. Recent history puts it beyond doubt that the first limb could plausibly have imposed a standard of belief. It did.

What was the rationale for the change? The decisive argument in *G* was that the majority in *Caldwell* had misinterpreted the Criminal Damage Act 1971 s1: references to 'recklessness' were continuous with the previous law's meaning of 'maliciously', and thus required actual advertence.[158] But, as Lord Bingham pointed out in *G*, the courts are slow to depart from (even mistaken) long-standing precedent supported by the highest courts.[159] The majority in *G* therefore relied on further arguments to justify departing from *Caldwell* and not requiring that D's belief be reasonable. These arguments focus primarily on incapable defendants.[160]

Immediately after *Caldwell* Glanville Williams seized on an ambiguity in Lord Diplock's judgment to argue that recklessness required that the risk would have been obvious *to the defendant* 'if he had thought about it'.[161] The effect of this interpretation would be to exclude from the scope of recklessness those *incapable* of understanding the risk. Williams was concerned that inadvertence due to mental incapacity should not be considered reckless, citing as relevant incapacities schizophrenia, dementia, and mental subnormality.[162] The courts quickly weighed in. In *Elliot v C* (1983), a 14-year-old girl, in remedial class at school, after a sleepless night wandering outside, destroyed a shed by setting fire to white spirit.[163] It was held that she gave no thought to the risk that the shed could be destroyed by her actions, and that she would not have realised the risk even if she had thought about it. Goff LJ dismissed Williams' interpretation of *Caldwell* on textual grounds and found the girl reckless. But he

---

[157] In *Caldwell* (n 12) Lord Diplock (at 354) referred to the 'ordinary prudent individual', whose hypothetical advertence defines what counts as an 'obvious risk'. This inter-definition was not clear, leading Glanville Williams to suggest an alternative 'conditionally subjective' reading whereby the risk must have been obvious *to the defendant* if they had considered it. ('Recklessness Redefined' (1981) 40 Cam LJ 252, 268-272.) Williams' suggestion was rejected in *Elliot v C* (n 12) 945-6, which clarified that the risk must be obvious to the 'reasonably prudent person'.

[158] *G* (n 12) [28]-[29], [35].

[159] ibid [30].

[160] The arguments' generality also explain why *G*'s interpretation of recklessness, though focused on property damage, has been applied to most other recklessness offences.

[161] Williams, 'Recklessness Redefined' (n 157).

[162] ibid 270-271, citing *Stephenson* [1979] QB 695, where the court accounted for D's schizophrenia in determining his (non-)recklessness.

[163] *Elliot v C* (n 12).

stressed that this result was 'unjust' and 'inappropriate'.[164] The girl was not reckless per the 'ordinary meaning of the word'.[165] Presumably, it was unjust because she was incapable of adverting to the risk. These misgivings took centre stage in *G*. There two boys, of 11 and 12, left a lit fire under some bins. The fire spiralled out of control and destroyed property worth £1m. Lord Bingham refused to apply *Caldwell*'s test because of the 'obvious unfairness' of convicting the boys. He continued:

> It is neither moral nor just to convict a defendant (least of all a child) on the strength of what someone else would have apprehended if the defendant himself had no such apprehension.[166]

He declined the invitation to relativise the standard to that of 'normal reasonable children of the same age', partially because

> if the rule were modified in relation to children on grounds of their immaturity it would be anomalous if it were not also modified in relation to the mentally handicapped on grounds of their limited understanding.[167]

The common denominator between children and the mentally handicapped is presumably their shared inability to advert to relevant risks.[168] The unfairness of finding incapable defendants liable seems to be the primary rationale for omitting the standard of reasonableness for D's beliefs.[169]

---

[164] ibid 947

[165] ibid 949.

[166] *G* (n 12) [33].

[167] ibid [37].

[168] Lord Steyn ([54]) also highlighted the potential unfairness to 'an adult who suffers from a lack of mental capacity or a relevant personality disorder'.

[169] Simester et al, *Simester and Sullivan* (n 38) 152-3: '[E]specially where the defendant suffers from limitations of age, intellect, or the like, to apply the *Caldwell* standard would be manifestly unfair, since it could lead to the conviction of persons…who lacked even the *capacity* to appreciate [relevant] risks… [I]t seems harsh to treat as reckless those whose inadvertence was due to preoccupation or distraction… It is even more harsh to equate with advertent recklessness the actions of those who were *incapable* of perceiving the risk.' (Emphases in original, footnotes omitted).

In sum, the law used to require that D's beliefs be reasonable. Its rationale for omitting that standard was to account for incapable defendants. The first limb of recklessness is a counterfactual incapacity relativisation.[170]

4.2 Self-defence: belief in threat

Self-defence is a complete common law defence. It's available where D:

1) genuinely believed that they (or another) faced an imminent[171] threat, and
2) they responded to the threat using a reasonable degree of force[172]

The standard of 'reasonable force' isn't relativised.[173] (Probably).[174] But what about the first limb? Does the absence of a standard of belief imply counterfactual relativisation?

---

[170] Recklessness could be interpreted as single-limbed: that *D unreasonably took a risk*. Then the *Caldwell/G* dispute is about the interpretation of 'unreasonably': is it unreasonable to take a risk only if you advert to it, or possibly unreasonable even if you don't? On this framing, the dispute is about the actual standard of reasonableness, not about counterfactual standards in the absence of a standard. Thus it can't be about counterfactual relativisation. But, if anything, this strengthens my argument. For on this framing the issue is whether 'reasonably taking a risk' is a relativisation proper. That *G* says that advertence is required for reasonableness constitutes an implicit relativisation, as the reasonableness standard's advertence requirement accounts for D's incapacities. The fact that a subtle difference in framing switches our interpretation from counterfactual relativisation to a relativisation again demonstrates the importance of including counterfactual relativisations.

[171] *AG's Reference (No 2 of 1983)* [1984] QB 456.

[172] As clarified by the Criminal Justice and Immigration Act (CJIA) 2008 s76.

[173] *Martin* (n 93); *Canns* [2005] EWCA Crim 2264; CJIA 2008 ss76(1)(b), (5A), (6), (9). Confusingly, s76(7)(b) says that 'evidence of a person having only done what the person honestly and instinctively thought was necessary for a legitimate purpose constitutes strong evidence that only reasonable action was taken.' Several defendants with mental disorders which led to their responding with (objectively) unreasonable force claimed that the force used was nonetheless reasonable because they honestly and instinctively believed it to be reasonable. But the courts dismissed this argument by noting that the act only intended to clarify the common law (s76(9)), which did not relativise 'reasonable force': *Oye* [2013] EWCA Crim 1725 (schizophrenia); *William Press* [2013] EWCA Crim 1849 (PTSD). This may be incompatible with a culpability principle: if *Oye* and *Press* were genuinely non-culpable, it is inappropriate to use conviction to achieve the (sensible) policy aim of restraining their dangerous conduct: ch 1 §3.

[174] There are cases in which it would seem patently unfair not to relativise the standard of reasonable force. Imagine D suffers a motor disorder (such that his physical movements are very clumsy) or an intellectual disorder (such that he doesn't know his own strength), meaning he responds to a threat with (objectively) unreasonable force. Stipulate: D was *not* any more vulnerable because of his incapacities; he knew precisely the gravity of the threat; he knew what would constitute reasonable force; but he failed to use it. Legally, D has no defence. That seems unjust. Most likely, his case would not be prosecuted, or else a judge would bend her interpretation of the facts, or else a jury would acquit regardless. But if those subterfuges were avoided, it seems plausible that the standard *would* be relativised in these circumstances to avoid injustice. (Especially because the policy concerns driving *Oye* and *Press* (n 173), where the threat was non-existent, are not relevant).

English law required a standard of reasonable belief in a threat for over a hundred years.[175] In that, it mirrored the requirement for a reasonable belief in consent in sexual offences. But the Lords omitted the reasonableness standard for sexual offences in *Morgan* (1976). That approach was soon emulated for self-defence.[176] While the reasonableness standard was reintroduced by statute in sexual offences, it remains absent from self-defence. It is more than plausible that there *could* have been a standard of reasonable belief. There was.[177]

But what was the rationale for the change? Unfortunately, the cases offer little help. The main reason provided was simply to harmonise self-defence with sexual offences.[178] That could be harmonisation for harmony's sake. If so, that rationale no longer exists after the reintroduction of the reasonableness standard in sexual offences. More charitably, the point of harmonisation was to extend the rationale underlying the change in sexual offences. But *Morgan* omitted the reasonableness requirement largely on grounds of ordinary language: it was not 'rape' in ordinary terms, the majority thought, if the defendant genuinely believed the victim consented.[179] It's not clear if the same argument transfers to self-defence. The only hint of a deeper rationale is found in *Beckford*, where it was noted that

> Where there are no reasonable grounds to hold a belief it will surely only be in exceptional circumstances that a jury will conclude that such a belief was or might have been held.[180]

---

[175] *Foster's Case* (1825) 1 Lew 187; *Rose* (1884) 15 Cox CC 540; *Chisam* (1963) 47 Cr App R 130.

[176] *Williams (Gladstone)* (1983) 78 Cr App R 276; *Beckford v The Queen* [1988] AC 130; *Martin* (n 103)

[177] '[P]rior to… *Morgan*… the whole weight of authority supported the view that it was an essential element of self-defence…[the accused's belief in a threat] was based on reasonable grounds.' *Beckford* (n 176).

[178] *Williams (Gladstone)* (n 176) 280-281, citing agreement with *Kimber* [1982] 1 WLR 1118, 1122, which found 'difficulty in agreeing with…[the] reasoning' of *Albert v Lavin* [1982] AC 547, where it was argued that the law ought to be different between the two offences. The argument was framed in terms of whether an offence had been committed in the first place, based on the meaning of the word 'unlawful' (as in 'unlawful assault'). Lord Lane approvingly cited Kimber for the proposition that offences include the word 'unlawful' to avoid making 'social life…unbearable, because every touching [otherwise] would amount to a battery unless there was an evidential basis for a defence.' This implies that acting in self-defence is not to apply unlawful force, and therefore is not a defence but rather an absence of an offence. See too *Beckford* (n 176) 144-145.

[179] *DPP v Morgan* (n 12) (Lord Cross, 203 and Lord Hailsham, 209, 214). (The Sexual Offences Act 1956 offered no definition).

[180] *Beckford* (n 176) 145.

Now imagine there were *no* circumstances where a jury could conclude that an unreasonable belief was genuinely held. Then the distinction between requiring a reasonable belief and requiring a genuine belief would be without a difference. De facto, a reasonable belief would be required. The courts are usually pragmatically-minded. Why not close the gap? Precisely, it seems, to account for those 'exceptional circumstances' in which the two come apart. For any ordinary defendant, it will be hard to prove a belief is held if unreasonable, and hard to argue that a requirement of reasonable belief is inapt.[181] Even the majority in *Morgan* saw no moral objection to that.[182] But such a moral objection might arise if D was *incapable* of holding reasonable beliefs in the relevant circumstances. D's incapacities might be the kind of 'exceptional circumstances' *Beckford* had in mind. When *Martin* later referred to 'exceptional circumstances' to relativise the standard of reasonable force in self-defence, it was incapable, mentally disordered defendants who sought (albeit in vain) to rely on those words. *Beckford*'s 'exceptional circumstances' plausibly refers to the same kind of incapable defendant, and thus that the rationale for the absence of a standard is to account for incapacities.

This is speculative. There may be no good rationale for the omission of a requirement of reasonable belief in a threat.[183] Or the primary rationale may be to account for mistakes, *simpliciter*. But it is at least plausible that requiring only a genuine belief in a threat is a counterfactual incapacity relativisation.

4.3 Loss of control: fear of violence

Loss of control has two qualifying triggers. One trigger refers to things said or done of an extremely grave character that caused D to have a justifiable sense of being seriously

[181] In *Williams (Gladstone)* (n 176) V was restraining a young robber. D thought that V was simply attacking the robber, and subsequently assaulted V. This seems like a reasonable mistake (indeed, V falsely claimed he was a police officer, further raising suspicions). D's initial conviction implies that a jury thought his mistake was not reasonable. Either the stated facts were misleading, or that seems like a very harsh jury decision. If the jury decision was harsh, this (morally) supports allowing D's appeal (by lowering the standard). But if the jury was right and D acted unreasonably, his negligence would likely be sufficiently culpable to warrant conviction.

[182] *DPP v Morgan* (n 12) 203 (Lord Cross): 'there is nothing unreasonable in the law requiring a citizen to take reasonable care to ascertain the facts relevant to his avoiding doing a prohibited act… [I]t is only fair to the woman and not in the least unfair to the man that he should be under a duty to take reasonable care to ascertain that she is consenting'. Lord Cross simply thought that this morally acceptable way for the law to be was not how the law was, given that it criminalised 'rape' and not 'non-consensual sex'.

[183] Though if a standard of reasonable belief was imposed it surely ought to be relativised.

wronged.[184] (The 'anger' trigger.)[185] But D's incapacities are merely incidentally relevant to the 'extreme gravity' standard[186] and the 'justifiability' standard is not relativised.[187] The other qualifying trigger is D's 'fear of serious violence' from V.[188] (The 'fear' trigger.) The explanatory notes explain that,

> As in the complete defence of self-defence, [D's fear of serious violence] will be a subjective test and the defendant will need to show that he or she lost self-control because of a genuine fear of serious violence, whether or not the fear was in fact reasonable.[189]

Like self-defence, the fear trigger is plausibly a counterfactual incapacity relativisation.

There's no precedent for a standard-setting version because the fear trigger was only introduced in 2009.[190] But it's revealing that the fear trigger didn't follow the anger trigger's demand that D's emotions be justifiable. That was because the fear trigger was conceived as an extension to the law of self-defence. It was motivated by concerns that abused children and women (and householders) may face disproportionately stronger aggressors. These victims may not have the means to use merely reasonable force or to wait until a threat was imminent.[191] Thus self-defence would not be available. That meant they faced the galling

---

[184] CJA 2009 s55(4).

[185] The trigger labels are from the Law Commission, *Partial Defences to Murder* (n 80) [3.38], noted by Alan Norrie, 'The Coroners and Justice Act 2009: partial defences to murder (1) Loss of Control' [2010] Crim LR 275.

[186] See §1.3.

[187] Explanatory notes to CJA 2009 s55, para 346.

[188] CJA 2009 s55(3).

[189] Explanatory notes to CJA 2009 s55, para 345. This means that a genuine fear is sufficient, even if mistaken or otherwise baseless. Curiously, this option is apparently foreclosed on the anger trigger, as s55(4) refers to 'things done or said' by another and not to D's perceptions of something done or said. This contrasts with the Law Commission's original proposal, which specifically permitted allowance for mistaken beliefs in (angering) provocations. (Albeit 'intelligible' mistakes). Law Commission, *Partial Defences to Murder* (n 80) [3.153]-[3.160]. However, the Law Commission's actual recommendations ([3.168]), which were largely implemented in the 2009 Act, made no allowance for D's (mistaken) perceptions. (Nor do the explanatory notes make any such allowance).

[190] The common law provocation defence (plus the Homicide Act 1957 s3) corresponded mainly to the anger trigger. Lord Hoffman in *Smith (Morgan)* [2001] 1 AC 146, 168 did say 'the law now recognises that the emotions which may cause loss of self-control are not confined to anger but may include fear and despair' (as noted by the Law Commission, *Partial Defences to Murder* (n 80) [3.85]), but without elaborating or citing any authorities. Arguably several 'battered women' cases fall under this heading, but at any rate the formulation in the 2009 Act is new.

[191] Law Commission, *Partial Defences to Murder* (n 80) [4.17].

choice of choosing between, on the one hand, murder and a life sentence (for using unreasonable force), or, on the other, victimisation.[192] The fear trigger provides a partial defence, and therefore flexible sentencing, to avoid that dilemma.[193] This implies counterfactual incapacity relativisation in two ways. First, the rationale for the trigger was to account for defendants who are comparatively less able than their aggressors, and especially cases of 'slow burn' domestic violence where defendants' abuse results in an inability to seek proportionate escape or defence.[194] Second, the fear trigger is essentially an extension of self-defence. The rationale for its absence of a standard is reasonably identical to that of self-defence. I concluded above, if speculatively, that this was to account for incapacities. Thus, insofar as self-defence constitutes counterfactual incapacity relativisation, so too does the fear trigger of loss of control.

4.4 Dishonesty: beliefs in circumstances

The mens rea for theft and fraud (inter alia) is dishonesty.[195] The test for dishonesty asks:

1) What did D believe about the circumstances?
2) Given this belief, was D's conduct dishonest according to the standards of ordinary decent people?[196]

In theft, it is not dishonest to take property if one honestly believes that one has a legal right to it, or that the owner would consent, or that the owner can't be discovered by reasonable steps.[197] Some defendants may mistakenly believe in circumstances that would make the appropriation honest (or consented to, etc.) due to incapacities. Is this merely incidental? Or does the absence of a standard for D's belief constitute counterfactual relativisation?

---

[192] ibid [4.18]-[4.24].

[193] And, practically, it allows defendants to run both self-defence and provocation defences rather than choosing between them and risking a murder conviction: ibid [3.89]-[3.91].

[194] John Gardner objects that battered women cases should not be understood in incapacity-based terms: 'Provocation and Pluralism' in *Offences and Defences* (n 56), discussed ch 3.

[195] Theft Act 1968 s1, s13, 17, 20, 22, 24A; Fraud Act 2006 s2(1), s3, 4, 11.

[196] Paraphrased from *Ivey v Genting Casinos* [2017] UKSC 67 [74].

[197] Theft Act 1968 s2(1).

There's no English precedent in favour of requiring reasonable beliefs, at least since the 19th century.[198] Given the settled nature of the law, the issue rarely arises in court.[199] But things could have been different. The American Model Penal Code demands a reasonable belief that the owner would consent.[200] And several peers endorsed (unsuccessful) amendments to the Theft Bill which would have required a reasonable belief in lawful authority.[201] A reasonableness standard was a live and plausible option, consciously rejected.

The peers proposing a reasonableness requirement all raised examples of unreasonable defendants: someone irrationally opposed to nude paintings; a pregnant mistress demanding excessive support with the 'venom of a jaundiced mind'; someone with 'a slanted view, an extreme view'; someone using 'violent, most objectionable means…[but who] because of his particular mentality [believed] that they were proper means.'[202] All were worried that such defendants could escape liability because of their genuine beliefs (paralleling concerns in self-defence and sexual offences). But the opposing Lords' response was that it is simply not dishonest to take something if one genuinely believes in facts that would make it honest, including the owners' consent, etc.[203] Thus dishonesty offences must ask only what D genuinely believed. In addition, they thought the worries raised were not practically problematic, as 'the taker is unlikely to believe that he is lawfully as distinct from morally justified.'[204] Reading between the lines, the issue seems to be the same as in self-defence, discussed above. If juries would *never* consider unreasonable beliefs to be genuinely held, then

---

[198] *Bernhard* [1938] 2 KB 264 cites various authorities on the point, including East's *Pleas of the Crown*, Stephen's *History of the Criminal Law of England*, and Halsbury's *Laws of England*, as well as *Leppard* (1864) 4 F&F 51, *Wade* (1869) 11 Cox CC 549, and *Clayton* (1920) 15 Cr App R 45. (Albeit those earlier cases tended to be brief and ambiguous about the precise ratio).

[199] An exception is *Holden* [1991] Crim LR 478, which simply confirmed the standard position.

[200] ALI, *Model Penal Code* (n 35) §223(3)(c). Curiously, no reasonableness requirement is added to the belief that the property was another's, or for a claim of right.

[201] Similar amendments were proposed to require a reasonable belief that demands were legitimate in blackmail. HL Deb 8 April 1968, vol 291, cols 100-137. For related discussion, see HL Deb 5 March 1968, vol 289, col 1302; HL Deb 7 March 1968, vol 289, cols 1493-1511.

[202] HL Deb 8 April 1968, vol 291, cols 100; 120-130.

[203] HL Deb 8 April 1968, vol 291, cols 131-135 (Lord Morris and the Lord Chancellor). The Government closely followed the 'subjective' approach of the Criminal Law Revision Committee's Eighth Report.

[204] HL Deb 8 April 1968, vol 291, col 100 (Lord Stonham).

the test would de facto require reasonable belief.[205] The continuing emphasis on defendants' genuine beliefs must be to catch precisely the unusual cases where unreasonable beliefs are nonetheless, with good evidence, found to be genuinely held. The most obvious cases to fall under these exceptional circumstances will be instances of disordered or incapable defendants.

Once again, this is a tentative conclusion. Little evidence speaks to the rationale for the absence of a standard of belief. But the evidence there hints—I put it no higher—that the test of genuine belief in dishonesty offences may constitute a counterfactual incapacity relativisation.

Conclusion

My main aim has been to vindicate the taxonomy of incapacity doctrines, incapacity relativisations, and counterfactual incapacity relativisations. I think these categories capture the variety of means by which the law accounts for incapacities more precisely than the more common language of 'objective' and 'subjective' tests while preserving a useful analytic generality. The doctrines surveyed above do not exhaust their categories. They are only a sample. I don't expect every criminal lawyer to agree that all the doctrines surveyed count as incapacity rules. The strength of the case for insanity and diminished responsibility is matched by the limitations of the case for self-defence and theft. But there is at least a case to be made. Finally, I mentioned that this is not a philosophical *defence* of incapacity rules. That is for the next chapter. But I do think that the breadth and persistence of incapacity rules provides at least prima facie evidence that they are warranted, and that the argumentative burden should fall on those who would abandon or reconceptualise them.

[205] A point made by Viscount Colville, HL Deb 8 April 1968, vol 291, cols 126-127, who worried about a 'genuinely conscientious' jury who would apply the law as written and acquit defendants due to their extreme beliefs.

3 Incapacity Rules Defended

The last chapter surveyed the incapacity rules. This chapter defends them. I will assume that being incapable, in some ways, is morally exculpatory. I'll defend that assumption in chapter 5. But it doesn't follow automatically that the criminal law ought to account for morally exculpatory incapacities via its incapacity rules. Several objections are available to block that inference, objections we'll consider over this chapter and the next.

John Gardner accepts the propriety of *some* incapacity rules. But his work raises a series of objections as to the wide scope of my taxonomy. He suggests that this wide scope is in some ways disrespectful (§1), that it is incoherent to relativise reasonableness standards to account for incapacities (§2), and that some of my alleged incapacity rules are better thought of as accounting not for our varying capacities but rather our varying roles (§3). I'll consider each objection in turn.

1 Respect

If my taxonomy is right, the law accounts for our incapacities in many more ways than is usually thought. This might be regrettable. Indeed, it might be so regrettable as to undermine my taxonomy. This objection is pressed in various ways by John Gardner. He claims that incapacity-based explanations of conduct are demeaning and disrespectful. An expansive taxonomy of incapacity rules thus implies an expanded range of demeaning and disrespectful treatment by the criminal law. That doesn't make my taxonomy wrong. But it does imply that my taxonomy is pessimistic. We might then have reason to prefer alternative, more optimistic, explanations for the wide range of (apparently) incapacity-based elements found in the criminal law.[1] We'll come to Gardner's rival explanation later. But is he right that my taxonomy bears the burden of pessimism?

[1] This preference for relative optimism can be explained in various ways: (1) as a general criterion of theory selection (albeit less weighty than accuracy, truth, etc), at least as an intertheoretic tie breaker; (2) as an epistemic criterion of theory selection under uncertainty; (3) as a criterion of legal theory selection, including eg as a criterion of fit with legal sources. This is not the place to argue for these theses, and

Reconstructed, Gardner's argument is something like this:

1) We necessarily 'aim at excellence in rationality'[2]
2) Incapacity-based explanations are non-rational explanations
3) Offering non-rational explanations is to fall short of our aim of rationality
4) Falling short of our aim of rationality is demeaning
5) ∴ Offering non-rational explanations is demeaning

Let's take the premises in turn.

Premise (1) is intuitively appealing.[3] Gardner understands excellence in rationality to comprise two aims: (a) acting for adequate reasons and (b) explaining ourselves in terms of those reasons.[4] He combines these in the notion of 'offering' an explanation which features in premise (3). I'll return to this in a moment.

Premise (2) is similarly appealing. To explain conduct in terms of incapacities is not to explain it in terms of acting for reasons, but rather based on limitations to acting for reasons.

Things get trickier with premise (3). Recall:

3) Offering non-rational explanations is to fall short of our aim of rationality

Gardner illustrates the idea with the well-known 'battered women' cases. Several women suffered abuse from their partners, 'snapped', and killed their abuser. They were charged with murder. Self-defence is rarely available in such cases: either lethal force would be disproportionate to the abuse suffered, or else the killing would not be in response to a fear of imminent attack. The defendants were not insane. They therefore had only two ways to avoid murder convictions (and mandatory life sentences). They had to choose between pleading provocation (now loss of control) and pleading diminished responsibility. On Gardner's view, they clearly ought to prefer provocation. Provocation entails that D had 'adequate reasons to get angry to the point at which [they] killed.'[5] That is, that D was *reasonably* angry. A rational explanation. By contrast, diminished responsibility is an incapacity-based

clearly objections are possible (eg perhaps a preference for pessimism would spur us to action). But the optimism preference is one I find plausible and at any rate it favours Gardner's critiques.

[2] John Gardner, 'The Mark of Responsibility' in *Offences and Defences* (Oxford 2007) 178.

[3] But cf Niko Kolodny, 'Why Be Rational?' (2005) 114 Mind 509; John Broome, *Rationality Through Reasoning* (Wiley Blackwell 2013).

[4] Gardner, 'The Mark of Responsibility' (n 2) 183.

[5] ibid 180.

explanation: it entails that one lacked the capacity to conform to reason; that one was unreasonable. It is better to be reasonable than not, and therefore better not to offer an incapacity-based explanation.

But that all comes with an important caveat. It is better to plead provocation than to plead diminished responsibility *only if that explanation is true*. As Gardner puts it, offering a rational explanation is to offer justifications and excuses, and doing that

> implies an ability to *have* a justification or excuse[;]… to give a rational explanation for one's actions without giving one's actions any rational explanation that they didn't actually have, i.e. without inventing reasons for what one did…[6]

It would have been better for the women in those cases to have *had* a rational explanation. Then they could truthfully have pleaded provocation. But pleading provocation if they lacked such an explanation would be, as Gardner points out, mere rationalization.[7] *Having* a non-rational explanation is to have fallen short of our aim of rationality. But, contra (3), *offering* a non-rational explanation is sometimes—when it is true—the only way *not* to fall short of that aim.[8]

Now take (4). Recall:

4) Falling short of our aim of rationality is demeaning

The demeaning nature of non-rational explanations might mean that we had better avoid them *even if accurate*. Especially, perhaps, where both rational and non-rational explanations are on

---

[6] ibid 183.

[7] ibid. Gardner does say that 'Since all rational beings want to assert their basic responsibility, all else being equal they cannot but welcome whatever contributes to that assertion.' ibid 192. But all else isn't equal. As Gardner himself puts it, we are 'aiming at successful understanding of the world…and not at mutual persuasion.' ibid 188. See too Victor Tadros, *Criminal Responsibility* (Oxford 2005) 145-146.

[8] Gardner claims that capacity-responsibility ('basic responsibility') is

> an ability which straddles the temporal gap between the wrong or mistake and the trial or recrimination, and which also straddles the conceptual gap between the ability to respond to reasons in what one originally does or thinks or feels, etc., and the ability to use those same reasons in explaining what one did or thought or felt. [It is] a compound—*not a mixture but a compound*—of our ability to use reasons in acting, thinking, choosing, wanting, etc. and our ability to use those reasons *again* in giving an account of whatever it was we did, thought, chose, wanted, etc., and in that sense, as rational beings, giving an account of ourselves. (Gardner, 'The Mark of Responsibility' (n 2) 184-185)

But this can't be correct. The law distinguishes insanity (lack of capacity at the time of the act) from unfitness to plead (lack of capacity at the time of the trial) precisely because it is helpful to ask *at which time* the defendant had capacity-responsibility. Forcing capacity-responsibility to straddle act and explanation simply results in confusion.

the table, or where there is uncertainty as to which was operative. (Did I really lose self-control, or did I act out of reasonable anger?) Gardner's view about non-rational explanations is asserted in no uncertain terms. A woman who pleads provocation leaves her 'head held high as a rational being', while one who pleads diminished responsibility 'demeans herself as a rational being'.[9] Only the provocation plea maintains her 'self-respect' in offering an 'intelligible rational account of herself.'[10] Indeed, 'One's responsibility is closely bound up with one's humanity, and to have it called into question is, with the best will in the world, degrading'.[11] The intuition is easy to grasp. Imagine that you raised an objection to my argument above, and I replied by saying that you were incapable of understanding it. You might feel patronised, condescended, demeaned.[12]

But not every non-rational explanation is demeaning. Imagine I'm blown into you by the wind. It would be absurd and unnecessary for me to struggle for a rational explanation for my non-rational conduct. The same applies if my knocking into you was non-rationally explained by my being woozy from general anaesthesia. Not all non-rational explanations are degrading.[13] So what makes non-rational explanations demeaning?

I think that intuition arises only if such an explanation is also *inapt* or used *in preference* to rational explanations. That was the situation faced by the defendants in the battered women cases. They could truthfully point out that their anger was reasonable. It was objectionable that they were being forced, by artificial limitations on the old provocation defence, into an inapt non-rational explanation. That is why their plight was objectionable.[14] Analogously, it would be

---

[9] ibid 180.

[10] John Gardner, 'The Gist of Excuses' in *Offences and Defences* (n 2) 133.

[11] John Gardner, 'In Defence of Defences' in *Offences and Defences* (n 2) 85.

[12] This aversion is familiar in many domains. It is *ad hominem* to attack a speaker's abilities rather than their argument. We might have *standing* to criticise others' arguments but not capacities. We might demand that we have a right to *punishment*, at least where the alternative is the condescension of treatment or the Caesarean arrogance of clemency. (GWF Hegel, *Elements of the Philosophy of Right* (1820, Nisbet trs, Wood ed CUP 1991) §100; Mary Beard, *SPQR* (Norton 2015) 223). Cf Jesus' remark upon being condemned: 'Father, forgive them; for they know not what they do'. King James Bible, Luke 23:34.

[13] Alan Bogg and John Stanton-Ife make this point in the context of consent to exploitative transactions: 'It is not clear how crediting [incapable] individuals…with a level of agency that they cannot reach would further respect for persons. A system of criminal law sensitive to such vulnerability is more likely overall to secure respect for persons than a system that is not.' 'Protecting the Vulnerable: Legality, Harm and Theft' (2003) 23 Legal Studies 402, 417.

[14] I therefore find Gardner's argument for liberalising provocation compelling, albeit on narrower grounds. That argument was taken up by the Law Commission, *Partial Defences to Murder* (Final Report, 2004), and eventually Parliament: Coroners and Justice Act 2009 ss54-56.

demeaning for me to claim that you were incapable of understanding my argument *if that was false*, but not if it was true. (Because, for instance, I wrote it in Swahili). Absent those qualifiers, however, there is nothing automatically demeaning about relying on non-rational explanations of our past conduct. To offer an intelligible rational account of ourselves now is to offer an *accurate* account of our past conduct. If that involves incapacity-based explanations of our past conduct, then so be it. We are not thereby demeaned.

My generous taxonomy of incapacity-based explanations in the criminal law would only bear the burden of pessimism if it also undermined or undercut more apt rational explanations. Expanding the legally-recognised set of incapacity-based explanations is not pessimistic in and of itself.

2 Reasonableness

Not all non-rational explanations are demeaning. But some are. They might be demeaning, and always inferior, if one has the choice between non-rational and rational explanations for one's conduct. Gardner's choice of example, the battered women cases, are persuasive precisely because of the choice available between provocation and diminished responsibility. But I claimed that provocation *itself* might be an incapacity doctrine, and that the fear trigger counted as a counterfactual relativisation. Further, I claimed that many doctrines which deploy reasonableness tests relativise that standard of reasonableness to account for defendants' incapacities. If I'm right, it's hard to hide from (partially) non-rational explanations. My taxonomy might then remain unacceptably pessimistic for seeing non-rational explanations wherever we turn. That is, they might undermine or undercut more apt rational explanations.

Gardner and Timothy Macklem raise this concern for provocation. They claim that relativisation makes 'provocation itself the very defence of mental abnormality that self-respecting defendants would rather not plead.'[15] They worry that to relativise reasonableness standards is to cast aspersions on *all* those who rely on them, including the fully capable. Self-respect demands a 'rigorously objective' standard, not a relativised standard.[16] The objection

[15] John Gardner and Timothy Macklem, 'Compassion without Respect? Nine Fallacies in *R v Smith*' [2001] Crim LR 623, 630 and 627.

[16] ibid 627.

can be cast in terms of fair labelling: relativised labels connote unflattering incapacities where none may be present.[17]

This version of the concern is overwrought. Consider self-defence. There is a world of difference between D1 fighting off armed robbers and D2 hitting V in the mistaken belief that V was such an armed robber.[18] Perhaps English criminal law ought to follow other jurisdictions in recognising that difference in its labels.[19] But does the law's failure to mark that distinction cast aspersions on D1? Does D1's self-respect demand that their conduct be labelled differently from that of D2? I doubt it. To demand that every element of every offence and defence ought to track every morally salient distinction is to demand too much of the law.[20] I agree entirely that crime labels ought to track culpability—I spent chapter 1 arguing as much. And there would likely be some minor upside to excising unflattering relativisations for those who need not rely on them.[21] But there would also be the large downside of failing to account for genuinely exculpatory considerations for everyone else.

Gardner's objection to relativising reasonableness standards focuses on that final point. For he denies that there is such an exculpatory upside to the practice.[22] He claims that to relativise reasonableness standards in the name of exculpation is not merely regrettable but positively

[17] Compare the objection to affirmative action that lowering standards will hurt beneficiaries who would meet the original standard.

[18] The kind of difference that Gardner has emphasised more than most: blameless wrongdoing still gives rise to 'outcome responsibility', moral duties to apologise, civil duties of restoration, etc. See John Gardner, 'Obligations and Outcomes in the Law of Tort' in Peter Cane and John Gardner (eds), *Relating to Responsibility* (Hart 2001), building on Tony Honoré, 'Responsibility and Luck' in his *Responsibility and Fault* (Hart 1999).

[19] Ie formally distinguishing between excusatory mistaken self-defence and justificatory self-defence.

[20] Gardner and Macklem could respond that they don't require this moral correspondence to capture all morally salient distinctions, but only those distinctions between rational and non-rational explanations. But the question then is why *this* is the most important distinction to track. There will be more moral variance *within* the sets of rational and non-rational explanations than between them.

[21] Though defendants can still explain themselves in 'rigorously objective' terms *outside* of the law. Gardner and Macklem claim that 'Systematically to bring the criminal law's standards of judgment down to meet people's incapacities…denies [defendants]... the fully human measure by which to account for themselves and hold themselves out for judgment.' But this is true only if we restrict our mode of accounting to legal categories. Gardner and Macklem, 'Nine Fallacies' (n 15) 627

[22] Many others claim, with Marcia Baron, that the reasonableness 'standard is supposed to be invariant, not to be relativized to the defendant'. (Marcia Baron, 'The Standard of the Reasonable Person' in RA Duff et al (eds), *The Structures of the Criminal Law* (Oxford 2011) 34). But I focus on Gardner's criticism as he explains the objection in its strongest form: that such a standard is incoherent.

confused. That's because to relativise reasonableness is to set a standard of *unreasonable* reasonableness. And that is a contradiction in terms.[23] Is he right?[24]

Once again, I doubt it. As we saw in the last chapter, the law routinely relativises its standards, including the reasonableness standard, to account for age and sex. The reason for doing so is perfectly intuitive: what counts as reasonable for D1 might not be reasonable for D2. Say I call you a 'fat head'. Getting angry at my childish insult is not reasonable; pity is more appropriate.[25] By contrast, even a self-respecting 10-year-old might find that such an insult crosses a line. It's perfectly understandable to say that such a child's anger might be reasonable. And it is reasonable because reasonable *for a child*.

Even if we accept that intuition, we might still worry that relativising reasonableness to a constituency leaves open the door to prejudice and discrimination.[26] If we ask what's reasonable *for a child*, why not *for a girl*? Thus relativised, juries might be tempted to answer that question by imposing a higher standard for girls than for boys (who, after all, 'will be boys').[27] Relativisation then becomes the handmaiden of discriminatory stereotyping. But, as Marcia Baron points out, this is a problem of applying the rule rather than with the rule itself.[28] If there is no defensible distinction to be drawn between a girl and boy's reasonableness, then there is no defensible ground for relativising to sex in such cases. Admittedly, it counts against any rule that it is easily misapplied. But that is not the case here: Parliament can easily remind

---

[23] This point is often repeated by judges and theorists. Eg *Lưc Thiet Thuan* [1997] AC 131 (PC); *Price* [2014] EWCA Crim 229 at 20: 'a defendant could never be found negligent… if the reasonable man were to be endowed with all the defendant's shortcomings, [as] any mistake which the defendant made could not be regarded as negligence.'; Mark Dsouza, 'Criminal Culpability after the Act' (2015) 26 Kings LJ 440, 442.

[24] Gardner has a separate objection that there is something incoherent about relativising any standard to account for *incapacities*. (Eg Gardner and Macklem, 'Nine Fallacies' (n 15) 626). I'll return to this (different) sceptical challenge in ch 6 §1.

[25] You might be reasonably angry at what my insult connotes, ie my infantilism or disrespect to you, but not at the insult per se.

[26] Eg per Sharon Byrd, 'On Getting the Reasonable Person out of the Courtroom' (2005) 2 Ohio State Journal of Criminal Law 571, via Baron, 'The Standard of the Reasonable Person' (n 22).

[27] The tort of negligence relativises its test of foreseeability in this way: 'whether an ordinarily prudent and reasonable 15 year old schoolgirl in the defendant's situation would have realised as much': *Mullin v Richards* [1998] 1 WLR 1304, 1308 (in which two girls fenced with plastic rulers, resulting in an eye injury). The criminal law's 'rough horseplay' exceptional allowance for consent as a defence to the causing of actual bodily harm might be used in a similarly gendered manner: *Jones* (1986) 83 Cr App R 375 (in which boys threw another into the air, dropped him, and caused a broken arm and spleen).

[28] Baron, 'The Standard of the Reasonable Person' (n 22) §5-6. Baron would, like Gardner, prefer a strictly nonrelativised standard of reasonableness to account for these worries. But the reply works equally well for a (partially) relativised standard.

judges, and judges juries, that some inferences ought not to be drawn from any particular attributes—say, that girls ought not to be held to higher standards than boys.[29] If the risk of discriminatory inferences is simply too high, this might justify ruling out certain relativisations altogether.[30]

Indeed, we can throw this objection back at Gardner's own conception of reasonableness. Reasonableness might be a standard of justification, but it isn't a standard of perfection. It doesn't (usually) demand supererogation.[31] But where, then, ought the line to be drawn between reasonable and unreasonable conduct? This line must account for human foibles. Reasonable force in self-defence cannot be precisely calibrated. One can have a reasonable belief in consent to sex even under the haze of alcohol.[32] Gardner and Macklem claim that the standard of reasonableness must be 'rigorously objective'. But what does this mean? Rigour and objectivity are not binary standards. One can be more or less careful, more or less aware of one's surroundings, more or less polite, respectful, attendant, and so on. The standard ought only to be as 'rigorous' as would be, well, reasonable in a broad swathe of circumstances, given a broad range of individuals. We can imagine that what is reasonable for a society of Superpeople would be correspondingly higher.[33] So we all accept—Gardner included—that the standard of reasonableness must be relativised to the *human* constituency. Otherwise, it could demand the supererogatory or the outright impossible.[34] But once we accept *this* relativisation, what compels us to relativise to but a *single* constituency? What does it matter whether Superman is surrounded by fellow Superpeople or mere humans? What is reasonable for Superman shouldn't depend on whether he happens to be on earth or Krypton. If so, we

[29] We might worry that judges and juries will ignore such directions, but it would be unduly pessimistic to suggest that no success is possible: see eg Berkeley Dietvorst and Uri Simonsohn, 'Intentionally "Biased": People Purposely Use To-Be-Ignored Information, But Can Be Persuaded Not To' (2018) J Exp Psych 1.

[30] As Parliament has done with sexual infidelity evidence in loss of control (Coroners and Justice Act 2009 s55(6)(c)), albeit rather ineptly (see *Clinton* [2012] EWCA Crim 2 [11]-[28]).

[31] Baron, 'The Standard of the Reasonable Person' (n 22) §6.

[32] Self-defence: Criminal Justice and Immigration Act 2008 s76(7)(a): 'a person acting for a legitimate purpose may not be able to weigh to a nicety the exact measure of any necessary action.' Consent to sex: *Bree* [2007] EWCA Crim 804.

[33] Let's stipulate that their superpowers extend beyond super-strength and laser vision (etc) to include super-virtues: courage, resilience, temperance, etc.

[34] A little discussed but important caveat is that the law does demand perfect consistency in living up to a less than perfect standard. That is, synchronic reasonableness is nigh impossible to achieve across a sufficient diachronic set. No driver of 40 years can have been reasonable in every moment spent on the road.

must accept at least two standards of reasonableness: super-reasonableness and human-reasonableness. But, in that case, why stop here exactly? Why not relativise to other constituencies, including to those of lower abilities?

These are not rhetorical questions. We might wish to impose a single standard to capture the value of uniformity, simplicity, equal treatment before the law, and so on. Indeed, one of the main functions of setting standards in the first place is to harmonise behaviour, to *set* expectations. Once expectations are set, it would be disruptive and confusing to recalibrate expectations when confronted with individual variance. We ought to stick to our guns and apply that standard of reasonableness, even to Superman, even though we can, extra-legally, expect much more from him. We ought to do this because, if we don't, we forfeit the unifying and harmonising benefits of standard-setting.[35]

We *might* want to do that. But we might not. Uniformity has its own costs. The most salient cost is that of failing to account for exculpatory specificities.[36] The question is not whether uniformity has value, but whether that value outweighs its costs. And I can see no reason for thinking that absolute uniformity strikes the proper balance. A standard of reasonableness that is relativised to (say) age need not destroy harmony or subvert expectations. It is both widely accepted and easily understood that children may be held to lesser standards across various domains: in customs, morals, and law. Indeed, to say that a single standard of reasonableness must apply invariantly to both adults and children would, if anything, surprise and subvert our pre-legal expectations. If we are concerned about uniformity out of concern for the rule of law, the law's moral legibility is just as important as its textual legibility.[37] A standard of reasonableness relativised to age then has the dual benefit not only of accounting for morally exculpatory factors but also of *upholding* clear expectations.

You might be wary of my examples. Children (and superpeople) are distinct and distinguishable from the rest of us. We therefore have clearly distinct and distinguishable expectations for them. But it doesn't follow that the same sort of variance will be so tolerable for those who differ in more subtle ways. There's certainly force in this point. But note that

---

[35] Consistently applying a standard despite sub-optimal individual circumstances has a well-known and fleshed-out strategic rationale in the theory of games. The locus classicus is Thomas Schelling, *The Strategy of Conflict* (1960, Harvard 1980) ch 5 (but cf ch 8). This theme is also developed by Frederick Schauer, *Playing by the Rules* (Oxford 1991) ch 7.

[36] As noted by Aristotle, *Nicomachean Ethics* (Bartlett and Collins trs, Chicago 2011) book V ch 10 (1137b ff).

[37] A point made by Gardner himself contra Andrew Ashworth: John Gardner, 'Rationality and the Rule of Law in Offences Against the Person' in *Offences and Defences* (n 2) 45.

my argument is for now simply responding to Gardner's claim that reasonableness must not be relativised at all. Accepting that we ought to relativise to age is sufficient to refute that claim. I haven't yet made any claim as to the extent of proper relativisation.

But perhaps there is a different problem with children. Some degree of unreasonableness (from an adult perspective) is perfectly ordinary and unremarkable for children. From this we might infer that to account for age is not really to relativise at all. As one Australian judge put it, the standard of (reasonable) care in tort cannot account for a child defendant's limitations that are 'personal to himself', but can if those limitations are 'characteristic of humanity at his stage of development and in that sense normal.'[38] (An approach blessed in English courts as having 'the advantage of obvious, indeed irrefutable, logic.'[39]) The claim is then that there is an invariant *human* standard, but humanity itself varies with age. I will have more say in the coming chapters as to whether an attribute's ordinariness or normality should make any moral difference.[40] But, for now, it suffices to point out that reliance on ordinariness is dangerously close to being question-begging. There are many statistically ordinary variants of humanity: in terms of raw numbers, the population with mental health disorders is about as large as that of those under 16.[41] Asking what counts as an 'ordinary' variant of humanity is not appreciably different from asking when relativisation is appropriate. And, to pre-empt those later discussions, it's hard to see what explains why we should care about ordinariness qua statistical prevalence, but not care about unusual candidate exculpatory factors like extreme circumstances or blindness.

Another reason to be wary of invariance is that different standards, including different standards of reasonableness, plausibly ought to differ according to their different contexts. What counts as a reasonable belief in another's consent to sex is plausibly different from that which would count as a reasonable belief in another's consent to surgery, and different again from what counts as reasonable force in self-defence. Perhaps these are simply incomparable contexts. But assuming that we can weigh them on a common scale of appropriate expectations, as Gardner does, must these different reasonablenesses require the same

[38] *McHale v Watson* (1966) 115 CLR 199, 213-214 (Kitto J).

[39] *Mullin v Richard* [1998] 1 WLR 1304, 1308 (Hutchison LJ).

[40] See ch 4 §2 and ch 5 §2.

[41] About 20% for each, at least in developed countries. Hannah Ritchie and Max Roser, 'Mental Health' (2019) <https://ourworldindata.org/mental-health> accessed 2019-04-18.

standard?[42] I doubt it. Plausibly a surgeon must do more to ascertain her patient's consent to having their leg amputated than she must do to ascertain her date's consent to having their leg touched. Now, perhaps this is just evidence of a *fixed* standard being applied to variable contexts: given the stakes, reasonableness demands more in the surgery case. But try reversing the severity of the cases. Our surgeon must remove a benign mole, and later ascertain if her date consents to sex. The surgery is trivial: the consequence of being mistaken would be the patient having a funny story for their friends. The sex is important: the consequence of being mistaken would be for the date to have been seriously sexually violated. Still—still—the standard of reasonableness to surgery demands more of her qua surgeon than the standard of reasonableness demands of her on her date. No consent form means no consent to surgery. But all sorts of mixed messages might not undermine a reasonable belief in sexual consent. If I'm right, the standard of reasonableness varies with context: not merely incidentally *according to* context but *altering to account for* that context.[43] It is *relativised.* If this is right, it is hard to see why the standard of reasonableness may be relativised to account for the context of the situation but may not be relativised to account for features of the situation that are internal to the defendant, to her attributes.[44]

Still, Gardner might worry, where does this leave us? What makes it the case that we may permissibly relativise to account for age but not mental disorders, or preferences, or habits, or weaknesses, or…? Won't all this relativisation mean that the 'useful generality of reasonableness suffers a death by a thousand cuts'?[45]

I don't see why it should. To accept that some relativisation is permissible is not to say that all is. As we saw above, we need to balance the benefits of uniformity (etc) with the costs of failing to account for exculpatory factors (and especially exculpatory incapacities). I said at the outset that the law frequently accounts for age and sex. Lawyers increasingly favour relativisation to intelligence too. Andrew Simester, for example, claims that 'we are not

[42] Gardner thinks of reasonableness as straightforward justification, applicable across different usages. John Gardner, 'The Many Faces of the Reasonable Person' (2015) 131 Law Quarterly Review 563.

[43] Marcia Baron accepts the first view but rejects the second. Baron, 'The Standard of the Reasonable Person' (n 22) 28.

[44] Kevin Tobia makes a similar claim as to how the standard of reasonableness might vary according to context: Kevin Tobia, 'How People Judge What is Reasonable' (2018) 70 Ala L Rev 293, 349. Hasan Dindjer has suggested to me that Gardner's suggestion for uniformity across contexts in Gardner, 'The Many Faces of the Reasonable Person' (n 42) fails to describe ('Wednesbury') unreasonableness in administrative law.

[45] Tobia, 'How People Judge What is Reasonable' (n 44) 348.

generally entitled to expect nonrelativised intellectual capacities from wrongdoers.'[46] The courts have accepted the power of this view in the law of duress, albeit grudgingly.[47] We can accept these relativisations, as the law and most theorists do, without thinking that this logically commits us to relativising to all attributes.

One final suggestion. Some have suggested that we ought to maintain an invariant standard of reasonableness but then to account for exculpatory factors at the excuse stage, as a defence.[48] Now, if this suggestion simply amounts to juggling around labels with no substantive consequences, then fine: juggle away. But two short responses. First, reasonableness often forms an element of defences: reasonable force in self-defence, reasonable firmness in duress, a reasonable (technically 'normal') degree of tolerance and self-restraint in the face of provocations. Shunting exculpatory factors into excuses to avoid relativising reasonableness standards doesn't work if those reasonableness standards are excuse elements. (They could be accounted for by other elements of a defence.) Second, one plausible substantive difference between offences and defences is that, roughly stated, offences put one on the hook of criminal liability whereas defences get one off it. Getting put on the hook, according to some theorists, entails having to answer for one's conduct at a trial. Getting off the hook requires that one offer an exculpatory answer. If this is approximately right, then it would be better in many cases for exculpatory incapacities to prevent defendants getting put on the hook (denying offending) rather than only getting them off it (offering a defence). A child might live up to (or exceed) expectations whilst still falling short of a standard of nonrelativised (adult) reasonableness. Why, then, must we go through the rigmarole of putting them on the hook only to take them off it afterwards?[49] This is not to say that this is *never* appropriate. It is only to say that it is not *always* appropriate.

In summary, relativising the standard of reasonableness is entirely commonplace in the law. It is intuitively appropriate to relativise to age and superpowers. And it is intuitively appropriate to relativise the standard according to context. This doesn't tell us the optimal extent of

[46] Simester's claim is notable in that he otherwise accepts Gardner's argument that 'our expectations of each defendant should *always* be relativised to the capacities of that particular defendant'. AP Simester, 'Wrongs and Reasons' (2009) 72 MLR 648, 667.

[47] *Antar* [2004] EWCA Crim 2708; cf *Bowen* [1996] 4 All ER 837, discussed in ch 2 §3.1.

[48] Baron, 'The Standard of the Reasonable Person' (n 22) 24-25.

[49] Luis Duarte d'Almeida, 'O Call Me Not to Justify the Wrong': Criminal Answerability and the Offence/Defence Distinction' (2012) 6 Crim Law and Philos 227.

appropriate relativisation. But it does imply that the answer is more than nothing. It is no flaw that my taxonomy of incapacity rules endorses incapacity relativisations.

3 Roles

I've argued that my taxonomy of incapacity rules is not unduly pessimistic and that reasonableness standards can be relativised. But a rival explanation for the (apparent) incapacity rules could yet prove superior. In particular, such a rival might avoid the difficulty, not yet tackled, of explaining exactly for which incapacities the law ought to account. I'll consider just one prominent rival: the idea that such rules can be explained by *role standards*.

I mentioned above that standards set expectations. This gave us one reason for favouring invariant standards: they set clearer expectations. But expectations also vary, and properly so. Some variations are incidental: they reflect changes in the underlying facts. But some variations are relativisations: the standards themselves differ. Expectations vary either way. The source of that variation needs explanation. John Gardner claims that it is our varying roles, not our varying capacities, which best explains the source of these varied expectations. He motivates the thought with the following case:

> *Checkpoint*: A car approaches a checkpoint in a troubled frontier zone. The soldier manning the checkpoint is young and inexperienced. The driver reaches down for her documents. Thinking that she's reaching for a weapon, the soldier panics and shoots her.

An incapacity-based view might endorse relativising the standard of reasonable force in self-defence to account for the soldier's inexperience, thus opening up a defence of (mistaken) self-defence.[50] Gardner claims that this would be inappropriate.[51] The fact that the solider was too jittery to make for a competent soldier discloses his unfitness for that role. This incapacity isn't something for which we ought to make allowances. Rather, the soldier's incapable failure to

[50] Alternatively, the incapacity-based view might defend the omission of a standard of reasonableness for the soldier's belief in a threat (as a counterfactual incapacity relativisation) whereas the role-based view might instead endorse requiring a standard of (role-based) reasonable belief in a threat. But only *might*, in both cases. To repeat, I haven't yet attempted to justify precisely for which incapacities (or roles) the law ought to account.

[51] Gardner, 'The Gist of Excuses' (n 2) 124-132. The case is modelled loosely on *Clegg* [1995] 1 All ER 334. In the real case the car was stolen and being joyridden.

meet his role standard is precisely why we ought to hold him liable.[52] We are entitled to demand a level of competence from the soldier commensurate with his role.

Role-based explanations cannot, and are not intended to, offer an alternative explanation to *all* the incapacity rules. The insanity defence and diminished responsibility, for example, are not amenable to a role-based explanation.[53] Gardner accepts that they are incapacity rules. But he can defend infancy, loss of control, and (especially) relativisations as means of accounting for varied roles rather than capacities.[54] When we say that D acted reasonably, we mean reasonably *given their role*.[55] Nor does this role-based explanation *contradict* incapacity-based explanations. The law could consistently account both for incapacities and for roles. These explanations could be exclusive: perhaps incapacities explain whatever roles cannot and vice versa. (Perhaps in *Checkpoint* the solider ought to be judged by role standards, but by incapacity-based explanations in other contexts). Alternatively, both explanations could be offered simultaneously. (Perhaps the soldier's liability in *Checkpoint* is assessed first by reference to his role, and second by reference to his incapacities, or vice versa.) The difference between role-based and incapacity-based explanations might only become apparent if the two explanations diverge in their prescribed outcome. The difference between role-based and incapacity-based explanations would then be one of emphasis. If so, Gardner's claim is that role-based explanations leave little unexplained for incapacities and/or ought to take precedence in the event of any conflicts.[56]

Gardner's argument, reconstructed:

6) We are morally excused only if our wrong actions do not reflect badly on us[57]

---

[52] ibid 124-132.

[53] As Gardner points out, the law shouldn't use role standards if nobody ought to be in that role, or if the law should not support the role's internal standards (eg because it entails being wronged or doing wrong). Nor too should the law use role standard if those standards are irrelevant to the wrongdoing. ibid 132.

[54] Gardner makes this case for loss of control (provocation) in many places. He has referred to infancy as an incapacity-doctrine (eg ibid 132) but in person (at my MPhil viva, 26 November 2015) has suggested that even infancy might be amenable to a roles-based explanation: that children simply ought to be less (eg) careful, more carefree, etc. Below I doubt this explanation.

[55] The claim is normative: that the law ought to account for roles. If the law doesn't, the role-standards view would criticise that.

[56] Gardner's main argument is negative: that incapacity-based standards are incoherent. I discussed this in the previous section and will return to the point more generally in the next chapter. This section is focused only on his positive defence of role standards.

[57] Gardner, 'The Gist of Excuses' (n 2) 130-131.

7) Wrong actions reflect badly on us only if we fall below the standard which can rightfully be expected of us (our normative expectations)[58]
8) Our normative expectations are determined primarily by the role(s) we occupy when wrongdoing (our role standards)[59]
9) We fulfil our role standards only by being fit for that role (which we are not if incapable of fulfilling its requirements)[60]
10) ∴ We are morally excused primarily only if we are fit for our role

Let's grant (6) and (7). Call any explanation of wrongdoing which negates culpability a 'moral excuse'.[61] The key claims are then (8) and (9). But why are our normative expectations determined primarily by our roles and not our capacities? And why can't we be 'fit' for our role despite being incapable of fulfilling all its rigours? Aside from general scepticism about capacity-based standards, scepticism we'll return to in later chapters, Gardner offers little by way of argument for these premises.

Start with (8):

8) Our normative expectations are determined primarily by the role(s) we occupy when wrongdoing (our role standards)

It's uncontroversial that *some* normative expectations are set by our roles. If my role as a salesman is to greet customers, we can rightly infer that I'm expected not to be unwarrantedly rude to customers. What is controversial is whether our roles are the *primary* determinant of our normative expectations. Might not role-based expectations pale in comparison to those set by my promises, by general morality, by my capacities? Gardner endorses the primacy of role-based expectations within an Aristotelian ethics in which roles set standards of virtue. But this commitment will not be shared by all. I for one doubt that our roles explain our normative profile—our rights, obligations, liabilities, etc—in a way that does not reduce to other more fundamental considerations. It seems plausible to me that teachers have rights over students, parents obligations to children, and citizens liabilities to the state only in virtue of contingent contractual, conventional, or otherwise contextual circumstances of each (type of) case. There are no parent/child normative relata that cannot be explained in terms of more fundamental

[58] ibid 124, 128-9.

[59] ibid 129-131

[60] ibid 130-131

[61] This is not Gardner's terminology: he calls incapacity-based 'moral excuses' 'exemptions'. I return to this point below.

or general reasons. (Eg that parents happen to be well placed to ensure the welfare of their children.) This is perhaps an unpopular view.[62] My point is only that the primacy of role-based expectations is not self-evident.

Gardner offers a negative argument against capacities-based normative expectations. This only supports (8) indirectly, by removing a rival. But, given it's the rival which I'm proposing, this seems good enough. Gardner's negative argument is reliance-based. The law solves coordination problems. It achieves the instrumental value of uniform expectations by setting role-based standards. And the instrumental value of uniformity 'militate[s] strongly against placing a capacity-based cap' on normative expectations.[63]

As I argued above, however, we care not about the value of uniformity on its own, but rather the *net* value of using any particular standard. We need some argument that the value of uniformity outweighs any disvalue associated with role-based standards. But that disvalue seem significant indeed. Consider the soldier in *Checkpoint* again. He was thrown into a tempest without the time to gain the fortitude or judgement of more senior personnel. Is he really to be judged on a par with them? Can we not make any allowance for his juniority? And doesn't this imply that we ought to account for his incapacities?

Gardner doesn't go so far as to rule out any leniency for the soldier. He apparently agrees with the intuition that rookie soldiers might be judged to a lower standard than more seasoned soldiers. But why? Because, Gardner claims, 'rookie soldier' is a distinct role from 'soldier', a role with lower built-in standards.[64] But why on earth should we want the role 'rookie soldier'? Would it not be better for soldiers to leave basic training with the fortitude and judgement of generals? There's no obvious reason to assign different roles to rookies *except to account for their incapacities*, to accept their limitations. If our individuation and selection of roles tracks and incorporates incapacity-based considerations, then the role-based explanation is not a rival to the incapacity-based explanation at all. It simply relabels it. Gardner's argument offers no persuasive support for the primacy of role-based normative expectations.

What about (9)?

---

[62] See, eg Nico Kolodny, 'Love as Valuing a Relationship' (2003) 112 Phil Rev 135; Nico Kolodny, 'Which Relationships Justify Partiality? The Case of Parents and Children' (2010) 38 PPA 37. For a recent, sceptical, overview, see Felix Koch, 'Skepticism About Special Obligations' (manuscript).

[63] Gardner, 'The Gist of Excuses' (n 2) 136.

[64] ibid 129. Gardner has suggested that this story generalises to children: perhaps they *ought* not to play with an adult level of care (see fn 54).

9) We fulfil our role standards only by being fit for that role (which we are not if incapable of fulfilling its requirements)

What does it take to fulfil our role-based normative expectations? Gardner claims that it requires *fitness* and that fitness for a role cannot make any allowance for incapacities to fulfil that role. But consider:

> *Seizure*: S is a brilliant surgeon. After many unblemished years at work, one day she suffers an unexpected seizure, resulting in a botched operation.
>
> *Dyslexia*: T is an excellent student, always working hard and demonstrating understanding in class. He struggles with written work due to severe dyslexia.

Is S fit for her role in spite of botching an operation? Is T fit for his role in spite of poor written work? This seems arguable. Yes: surgeons require steady hands and students require good writing. There is no role of 'surgeon-suffering-a-seizure' nor 'dyslexic-student'.[65] No: fitness for a role itself should account for (eg) effort as opposed simply to outcomes. Conclusion: unclear.

The more important point is what lesson we draw from such cases. If Gardner claimed that role standards were the *only* determinants of our normative expectations and that role-standards did not account for incapacities, then from the argument above he would have to conclude not only that S and T were unfit for their roles, but also that they lacked any moral excuse for their failures. That they were *culpable*. That can't be right. Assuming that S had no foreknowledge of the risk of such a seizure, we surely cannot blame her for botching the operation. Nor can we blame T for his poor writing insofar as it stems from his dyslexia. Gardner can accept this. He doesn't deny that some incapacity based moral excuses are available. But he tries to confine these to a very narrow category, a category he labels 'exemptions'. The problem, however, is that this category does not seem flexible enough to incorporate cases like *Dyslexia*. Recall the conclusion to the argument:

10) ∴ We are morally excused primarily only if we are fit for our role

[65] Gardner's discussion was mostly focused on standards of character, whereas seizures and dyslexia are not incapacities that touch upon our character. However, Gardner's argument was intended to apply more broadly than only to standards of character, including to standards of skill—including, then, the skills of surgery and of writing (ibid).

Naturally, the plausibility of the argument depends on the scope of that 'primarily'. But, as *Checkpoint* revealed, Gardner attempts to fit a great majority of moral excuses into the role-fitness framework. By contrast, cases of incapacity-based excuses, cases like *Dyslexia*, fall out of the picture, and potentially out of the scope of moral exculpation altogether. That, I think, is the wrong emphasis. Incapacities have a very significant role in establishing moral excuses. Living up to one's role seems, if anything, a minority mode of exculpation.[66]

Role-based normative expectations might play many roles in the law, but they don't seem well suited to adjudicating the availability or extent of moral excuses. They are no great rival to incapacity-based explanations in the law, and therefore present no great obstacle to my taxonomy of incapacity rules.

Conclusion

There are always more objections. I've limited my attention to those I find most plausible. But they also reveal the limitations to the argument so far. In particular, I've yet to respond to Gardner's general scepticism about incapacity-based standards. That is a task for chapter 6. Before that, however, I will consider a final kind of objection to the incapacity rules. It is that they identify relevant incapacities by the use of proxies—mental disorders, young age, etc—rather than asking directly and simply for the relevant incapacities themselves. Why should that be the case? This is the question for the next chapter.

[66] AP Simester gives the example of an athlete failing in their role due to a hamstring injury. His worry is that Gardner's idea of culpability comes too cheap. Gardner understands culpability to entail (a) a wrong, absent (b) exemption, (c) excuse, and (d) justification. Simester wants there to be something *active* in culpability, something like *choice*. Simester, 'Wrongs and Reasons' (n 46) 663-668. Jeremy Horder stresses the need for moral activity, versus passivity, in culpability: Jeremy Horder, *Excusing Crime* (Oxford 2004) ch 2.

4 Incapacities and Aetiologies

Chapters 2 and 3 surveyed and defended the incapacity rules. But that defence is not yet complete. For one thing, we still need to defend the premise that the relevant incapacities make defendants less culpable. That's the task for the next chapter. But, even granting that premise, it still doesn't follow that the law's incapacity rules are justified means of accounting for those morally exculpatory incapacities. There remains a gap between moral exculpation and legal exculpation. Bridging it is not entirely straightforward.

For one thing, some claim that we could account for those morally exculpatory incapacities perfectly well without any specific incapacity rules. The idea is that incapacities often result in defendants lacking the requisite mens rea for a crime. This allows incapable defendants to avoid liability, due to their incapacity, without the use of any dedicated incapacity rules. If that scenario generalises there is no need for any incapacity rules to bridge the gap between moral and legal exculpation. We could and should *abolish* incapacity rules (§1).

Another problem is that some incapacity rules only indirectly account for exculpatory incapacities. They specify *proxies* for incapacity-derived lowered culpability: young age, recognised mental illnesses, etc. But we need an additional argument to justify the use of these proxies. Gideon Yaffe has recently claimed that such arguments fail to justify leniency towards children. His argument might generalise to undercut the justification for these proxies (§2).

Finally, some incapacity rules are available only where the relevant incapacity derives from a specific *aetiology*. As with proxies, these aetiology requirements introduce a gap between the content of the incapacity rules and the underlying morally exculpatory incapacities. But, unlike proxies, these aetiology requirements apparently deny legal exculpation to provenly morally exculpated defendants. The effect seems to be to make the incapacity rules less accurate in picking out their target. What could justify that? (§3).

1 Abolition

Do we need incapacity rules? The case for abolition is simple. Imagine you fall and break a window. This opens you up to potential criminal damage liability. But you can explain! You're just clumsy! You didn't damage the window intentionally or recklessly. You lacked the mens rea for the offence. Thus, you're not criminally liable.

A different explanation: You weren't *able* to avoid breaking the window! You're only nine years old! Or: You were sleepwalking! That is: you were incapable in some way, and that incapacity meant you lacked the mens rea for the offence.

Just like clumsiness, your incapacity-based explanation means you won't be found criminally liable. The point? We don't need any dedicated legal rules to account for *clumsiness*. Likewise, then, perhaps we don't need any dedicated rules to account for incapacities.

Both explanations *deny offending*. Strictly speaking, it is for the prosecution to prove, to make a jury sure, that you possessed the requisite mens rea. Technically, then, you needn't offer any explanation for breaking the window. But let's be realistic. If you go flying through a window, with nothing to say for yourself, it won't take a jury long to think the worst. You're going to have to offer *some* explanation.[1] But—and this is the crucial point—it doesn't make any difference, plausibility aside, whether you cite clumsiness or lack of capacity. Both are cited to deny offending. It seems that we don't need any dedicated incapacity rules. The only rule we need is that defendants are allowed to cite their incapacities when denying apparent wrongdoing, just as they are allowed to cite their clumsiness.

At least some incapacity rules don't work like this. Some incapacity rules do not deny offending, but rather raise a supervening defence.[2] That's true of insanity, for example.[3] But the abolitionist proposal, unsurprisingly enough, is that these (variants of) incapacity rules ought to be abolished.[4]

The case for abolition doesn't rest only on the fact that such a system would be simpler. Simplicity ought not to be bought at the price of denying exculpation to those morally exculpated on the basis of incapacity. The abolitionist case is instead that incapacities are morally exculpatory only if they amount to denials of offending. Far from denying exculpation

[1] Adverse inferences can be drawn from silence: Criminal Justice and Public Order Act 1994 s35.

[2] Or, for (counterfactual) relativisations, an element within a defence.

[3] *Loake v CPS* [2017] EWHC 2855 (Admin).

[4] Considered by the Law Commission, *Criminal Liability: Insanity and Automatism* (Discussion Paper, 2013) ch 2.

to the morally exculpated, abolition would apportion liability exactly as it ought to be apportioned.

The abolitionist proposal wouldn't amount to much if we could just relabel what are currently defence elements as offence elements. Then the incapacity rules could operate more or less as they do now, simply under the label of denying offending. No: the abolitionist claim is more fundamental than this. The abolitionist claim, in the language of chapter 3, is that incapacities are *merely incidental* to culpability, and thus ought to be *merely incidentally relevant* to criminal liability. Incapacities might affect the facts, which in turn affect our normative position, our liability. But our incapacities do not *directly* alter our normative position.

The intuitive pull of the abolitionist proposal is best conveyed by two examples: automatism and the treatment of a defendants' beliefs.

Automatism is a specific explanation to deny offending. It's morally exculpatory only if it means one lacks the mens rea or actus reus of an offence.[5] Given that limited effect, why did I include automatism within my taxonomy of incapacity rules?[6] The reason I gave was that, unlike clumsiness, automatism is defined by distinct rules. Those rules are required to distinguish automatism (and an outright acquittal) from insanity (and the special verdict). That, I claimed, is why automatism counts as an incapacity doctrine whereas clumsiness does not, even if they both amount to denials of offending. But the abolitionist proposal undercuts this response. According to the abolitionist, we ought not to have *any* incapacity rules. Including insanity. That is: insanity should be legally relevant only insofar as it amounts to a denial of offending.[7] Then there would be no need to distinguish between automatism, insanity, and clumsiness as explanations for offending. There need not be any incapacity rules at all.

The second example relates to beliefs. I'm not criminally liable for attacking someone if I act in self-defence, assuming all the elements of that defence are present, regardless of why I

[5] JJ Child and Alan Reed, 'Automatism is never a defence' (2014) Northern Ireland Legal Quarterly 167.

[6] See ch 2 §2.4

[7] Stephen offered a doctrinal version of his argument. He claimed that all crimes require both *voluntariness* and *knowledge of the 'character' of the conduct* in question. On that view, the insanity defence is simply a negation of the latter. James Fitzjames Stephen, 'On what Principles ought the Law to deal with Questions of Responsibility and Mental Competence, in Civil and Criminal cases respectively?' in George Hastings (ed), *Transactions of the National Association for the Promotion of Social Science* (Longman 1865) 182. But Stephen's conditions are not required as a matter of English law. Alternatively, say that Andrew Ashworth is right to think that knowledge of the legal prohibition ought to be required for all offences, such that ignorance of legal wrong, no matter the aetiology, would exculpate. (To simplify his position). Andrew Ashworth, 'Ignorance of the Criminal Law, and Duties to Avoid It' (2011) 74 MLR 1.

believed I was facing an imminent threat. (Including, eg, that I was unable to distinguish genuine from delusional threats). I am criminally liable if I attack someone with intent, regardless of whether that intent was drunken. (Including, eg, that I was involuntarily intoxicated).[8] This lesson might generalise to the incapacity rules. Perhaps I am no more or less culpable for attacking someone with insane intent. Say I kill V after suffering delusions about V. The content of those delusions is all-important. If I believed that V was *attacking* me, that is mistaken self-defence. I am not criminally liable whatever that belief's origin. My incapacity is merely incidental.[9] Contrast: I attack V mistakenly believing that V belonged to *a certain ethnicity*. I am criminally liable whatever that belief's origin.[10] In both cases what matters is whether my conduct was justified or excused *given* my beliefs. The *origins* of my beliefs are beside the point, whether incapacity-based or otherwise.[11] Thus: the criminal law has no need for incapacity-specific rules: the general rules relating to beliefs sufficiently account for incapacities already.

These cases demonstrate the intuitive pull of the case for abolishing the incapacity rules. But they are not univocal in supporting that conclusion. Reconsider the analogy to intoxication. The thought is that intentions are no less culpable for being drunken.[12] But this is arguably too simplistic. Consider a defendant who foresees that he might form and act upon certain bad intentions if drunk, but when sober rejects, condemns, and foreswears such intentions. He vows never to drink and keeps that vow. He may lack Aristotelian virtue—the right desires and intentions don't flow easily—but he has no lack of Protestant grit. Now imagine his enemy spikes his drink, causing his intoxication, leading to his forming (and acting) on those rejected,

---

[8] *Kingston* [1995] 2 AC 355 HL.

[9] Stephen, 'Responsibility and Mental Competence' (n 7); Joseph Goldstein and Jay Katz, 'Abolish the "Insanity Defense" – Why Not?' (1963) 72 Yale Law Journal 853; Norval Morris, 'Psychiatry and the Dangerous Criminal' (1968) 41 So Cal L Rev 514; Norval Morris, 'The Criminal Responsibility of the Mentally Ill' (1982) 33 Syracuse L Rev 477.

[10] Assuming the other elements are present.

[11] Unfitness to plead is untouched by the objection for a different reason. The trial is the place for a defendant to answer for their (alleged) conduct, whether that answer admits or denies the offence elements. Unfitness at trial prevents a proper answering. That is true, and the doctrine ought to be retained, even if the objector is right that answers relying on incapacities only operate as denials of offending.

[12] *Sheehan and Moore* [1975] 1 WLR 739, 744. It's often claimed that intoxication is a 'defence', albeit only to crimes of 'specific intent'. But intoxication (where voluntary, for 'basic intent' crimes) allows the law to impute *absent* mens rea elements. It is, if anything, the opposite of a defence. It is simply concerned with making out (or making up) mens rea. Andrew Simester, 'Intoxication is Never a Defence' [2009] Crim LR 3. ('Appropriately', the objectors might add, and I would be inclined to agree). Arlie Loughnan, *Manifest Madness* (Oxford 2012) 30-31, ch 7, claims that intoxication is an ('imputation') 'mental incapacity doctrine', analogously to my incapacity rules.

condemned, and foresworn intentions. English criminal law holds him liable for his actions nonetheless, without any kind of 'defence'.[13] This is at least arguably too harsh.[14] D's case tells a more complicated story than the bare elements of the offence.

It would be harsher still for the law to ignore these kinds of contextual factors regarding insanity. Imagine a model of both Protestant and Aristotelian virtue. She succumbs to sudden and dramatic madness and comes to believes in malign conspiracies. As a result, she intentionally kills a notorious gang leader.[15] Is it plausible to maintain that her insane intent is intent nonetheless, no further questions asked? I find that hard to accept.[16] We should make some allowance for obviously disordered minds. Abolitionists tend to agree. Stephen grudgingly admits that such defendants are 'entitled to the benefit of a doubt', ie exculpation, notwithstanding the presence of mens rea.[17] Norval Morris, who denies that insanity ought to be a complete defence, still accepts that fairness requires *some* leniency for mentally incapable offenders.[18] This intuitive pull for exculpation despite all the offence elements being made out is stronger yet for infancy.[19] Children are not always morally exculpated because they deny wrongdoing. Sometimes it is simply their failure to appreciate the nature of their wrongdoing that cries out for exculpation.[20] As such, the intuitive case for abolitionism falls short of its

[13] *Kingston* (n 8).

[14] Victor Tadros, *Criminal Responsibility* (Oxford 2005) 319-320. Depending on how we interpret D's intoxicated state, this might violate a 'control principle' suggested by, eg Douglas Husak, 'Rethinking the Act Requirement' (2006) 28 Cardozo L Rev 2437, 2457-9.

[15] We could further stipulate that D believed facts such that V may (objectively, morally) justifiably have been killed (whether due to liability to lethal defensive harm (albeit not in imminent self-defence), or else simply to save many other lives). Even here English law offers no defence (self-defence or duress) but for insanity/diminished responsibility. Perhaps the law ought to alter those doctrines? I consider this move below.

[16] Stephen Morse, 'Crazy Reasons' (1999) 10 J Contemp Leg Iss 189, Michael Moore, *Law and Psychiatry* (Cambridge 1984) 223, endorsed by the Law Commission, *Insanity and Automatism* (n 4) [2.20].

[17] Stephen, 'Responsibility and Mental Competence' (n 7) 187. He wriggles this into an evidential claim: 'in so strange a state…*no one can be sure* as to the way in which he reasons.' (Emphasis added). But I take it that the intuition remains even if we stipulate away evidential worries.

[18] Norval Morris advocates a diminished responsibility defence on the grounds that it offers better procedural protections than civil commitments, and that it is the murder *sentence*, not conviction, that is the 'real evil'. In this he accepts the driving force of the argument that incapacities exculpate, only rejecting insanity due to highly contingent (and questionable) matters of institutional competence and procedural considerations regarding the protection of individual rights. Morris, 'The Criminal Responsibility of the Mentally Ill' (n 9) 513-4. Questionable, as Morris treats a stay in prison and a stay in hospital as equivalent; a matter of indifference for the defendant.

[19] Gideon Yaffe rejects this (see §2 below) but acknowledges that he's swimming against the tide.

[20] We could alternatively conceive the infancy defence as an aetiological proxy, a shortcut, for an absence of mens rea. I consider this in §3, below.

target. Not every (valid) incapacity-based explanation can be accounted for via the generally applicable rules.

But perhaps the abolitionist can accommodate these intuitions. Rather than introducing a complex suite of incapacity rules, she might reply, why not simply liberalise the generally applicable rules?[21] Rather than introducing a distinct insanity defence, for example, why not simply qualify the basic mens rea requirements for all offences to require that D understands the wrongfulness of their conduct? This is how Stephen thought insanity worked. Ashworth has suggested such a general reform.[22] We could similarly 'subjectivise' defences to judge defendants according to their actual beliefs, rather than imposing standards of reasonable belief, in order to account for incapacity-induced beliefs.[23] No incapacity rules required.

One difficulty with this reply is that it calls for widespread reform throughout the criminal law.[24] Doing so would incur transition costs and faces the challenge from conservatism. But it would also mean two big downsides relative to the incapacity rules as described by my taxonomy.

The first downside is that the abolition proposal demands uniformity. Consider the standard of reasonable belief in consent in sexual offences. For the abolitionist to rely only on generally applicable rules would mean either permitting no exceptions to this standard or else omitting the standard entirely. Omitting the standard would be to bring the law back to the position after *Morgan,* in which even an unreasonable but genuine belief in consent could shield defendants from conviction. That law had obvious costs. Most theorists consider reckless rape highly culpable. And, even if not, omitting any standard of belief strikes an undesirable balance between victims' and defendants' interests. The only alternative, on the abolitionist proposal, would be to impose an exceptionless and uniform standard of reasonable belief. But why demand this dichotomy? The law might be able to strike a better balance yet by accounting for some limited cases of non-culpable mistaken beliefs, such as incapacity-induced mistakes. This

[21] Eg, regarding the situation in fn 15 above, perhaps self-defence shouldn't require threats be imminent, or duress ought to judge defendants on the facts as they (even delusionally) believed them to be. That approach is advocated by, eg Christopher Slobogin, 'An End to Insanity: Recasting the Role of Mental Disability in Criminal Cases' (2000) 86 Virg LR 1199.

[22] See fn 7 above.

[23] Per Stephen's view of insanity, Gideon Yaffe has suggested that the generally applicable doctrines *already* obviate the need for a specific infancy defence (or other child-affecting doctrines) to account for children's exculpatory incapacities. I reject this in §2 below.

[24] My taxonomy of incapacity rules was precisely intended to capture how the law accounts for incapacities in these ways via its relativisations and counterfactual relativisations.

is possibly the law's approach to defendants incapable of understanding social cues. To strike this more nuanced balance requires a partially relativised standard of reasonableness. It requires an incapacity rule.

The second downside for the abolitionist's reforms is that the law would not distinguish between (on the one hand) counterfactual relativisations to account for certain attributes and (on the other) merely incidentally relevant attributes. If the abolitionist had their way, all attributes would be merely incidentally relevant. This would forgo any explanatory benefits of making the distinction. I suggested some of those benefits in the last chapter. I gave the example of the 'anger trigger' for loss of control. It is available where a threat is 'extremely grave'. What makes a threat count as 'extremely grave' is open-ended. Any of the defendants' attributes could affect the context of the threat, making it extremely grave in context: a provocation's gravity can be affected by D's relationship status, ethnicity, employment, incapacities, …ad infinitum. That differs from tests which can be affected in practice by any attribute, but which are designed specifically to account for some particular attribute(s), such as D's incapacities. While subtle, this difference is useful for theorists trying to understand the underlying structures of the law.

But the distinction between incidentally relevant attribute and counterfactual relativisations isn't only of theoretical interest. It's also useful for analogical reasoning. The courts omitted a standard of reasonableness for D's belief in a threat in self-defence to harmonise that law with the (old) law of sexual offences.[25] More recent cases have contrasted the two doctrines.[26] In other contexts the courts have refused to analogise.[27] When are such analogies appropriate? The courts have offered little guidance, other than (re)asserting the (dis)similarity of the doctrines under discussion. Distinguishing between counterfactual relativisations and merely incidentally relevant attributes allows us to discern at least one vector of (in)appropriate analogical reasoning. If a standard was omitted to account for incapacities there may be a

[25] See ch 2 §3.3.

[26] *B (MA)* [2013] EWCA Crim 3 [36]. The judgment also compares reasonableness in rape with reasonableness problems in provocation at [29] and [38] (coming out strongly in favour of the objective position in *AG for Jersey v Holley* [2005] UKPC 23 versus the relativising position in *DPP v Morgan* [1976] AC 182).

[27] Eg *Oye* [2013] EWCA Crim 1725 [43] re the subjective element of reasonableness in self-defence: 'It seems to us best not to seek to draw any comparisons with defences such as, for example, loss of control or duress where questions of honest, but mistaken, belief can also arise. As rightly noted in *Smith & Hogan's Criminal Law*, 13th ed (2011), p 383, there is, on the authorities, no clear coherence of approach in these areas. Indeed, the approach indicated in, for example, the highly complex provisions of sections 54–56 of the Coroners and Justice Act 2009 relating to loss of control would seem to indicate no particular parliamentary intention that a corresponding approach is designed to be adopted.'

principled basis (consistent with Parliamentary intent, etc) for adding in restrictions to the use of that standard, but only for non-incapacity-based attributes. In general, that a standard was omitted as a counterfactual incapacity relativisation tells us that incapacities, but not other attributes, ought to be accounted for when interpreting, revising, or expanding analogous doctrines. That the omission was merely incidentally relevant implies that the analogical reasoner has a freer hand. It is a mark against abolitionism that it would flatten this distinction.[28]

One final reply. I've defended the incapacity rules as necessary and sensitive means by which to account for exculpatory incapacities. The gist of the abolitionist complaint was that this is too generous; that incapacity-blind rules might be more appropriate. But the abolitionist could also argue that the current incapacity doctrines are too strict: that they *hinder* the law's ability to account for exculpatory incapacities by *preventing* defendants from benefiting from denials of offending. In that case we should abolish the incapacity doctrines in order to exculpate *more* generously.

Let me explain. The insanity defence is available where mens rea is present. But some insanity pleas amount to denials of mens rea.[29] The criminal law's 'one golden thread' is that the prosecution bears the duty to prove the defendant's guilt, including any mens rea elements.[30] The insanity defence subverts this burden. A defendant who denies mens rea due to a mental condition must rely on insanity *instead of* pleading for a pure acquittal. Given that defendants raise defences, why can't they simply refuse to plead insanity, anticipating the prosecution's failure to prove mens rea? In practice, *some* explanation for their conduct will be required to

[28] The objector might note that my responses here justify relativisations and counterfactual relativisations. What about the paradigmatic incapacity doctrines? In short, they can be justified in the same way as relativisations, for they are simply rules stripped of functions other than to relativise to incapacities. The objector might claim that they add unjustified complexity to the law. Perhaps the law doesn't exculpate those who form intentions after being involuntarily intoxicated—in spite of this being exculpatory—because *Kingston*-style cases are too rare to justify adding complexity to an already complex area of law. And perhaps the same is true of incapacity doctrines: perhaps cases of Aristotelian virtue, Protestant grit, and intentional criminal offending all coinciding are too rare to justify having distinct doctrines. But whether that is the case will depend on the weight of these respective reasons. In line with the culpability principle I defended in ch 2, complexity seems a reasonable price to avoid convicting the non-culpable. The incapacity doctrines are important final safeguards to avoid such outcomes and are relatively easily specified (as distinct and distinctive doctrines), and thus seem to warrant the added complexity.

[29] Some claim this is true for *all* cases relying on the nature limb: David Ormerod (ed), *Blackstone's Criminal Practice* (Oxford 2019) [A3.23]. But it seems that D might not know the nature of their act and yet still commit an offence of strict or absolute liability, to which insanity could yet be a defence, and thus the nature limb cannot merely be a denial of mens rea.

[30] *Woolmington v DPP* [1935] AC 462, 481.

avoid adverse inferences. But, additionally, the insanity defence can be put before the jury unilaterally by the judge[31] and a special verdict can be substituted by the Court of Appeal without the defendant's consent.[32] But the effect of this is to *deny* people who have not fulfilled the elements of an offence—innocent people—an outright acquittal.[33] In this sense insanity inculpates. It turns what would be an acquittal into something worse. It operates like intoxication, where proof of intoxication/insanity by the prosecution can suffice to ground liability *in lieu of* proving mens rea.[34]

Am I not overstating things? After all, the intoxication doctrine results in a guilty verdict, whereas the outcome of insanity is the special verdict. That is, at least formally, not an inculpatory conviction. But there is also the worrying possibility that a defendant might be denied *both* an acquittal *and* the special verdict if the insanity defence is raised but rejected. For example, if a defendant's condition is held not to amount to a 'defect of reason', or if their ignorance of wrong is not ignorance of *legal* wrong, the insanity defence will not be available. Once raised, however, an outright acquittal might be denied on the same factual basis. Indeed, this same worry might even affect automatism, which is limited to *total* destruction of voluntary control.[35] If a defendant denies mens rea on the basis of lacking control of their conduct which is deemed only *partial*, it seems possible that this denial might be rejected for falling outside the scope of automatism, and thus a guilty verdict reached despite an absence of mens rea.[36]

---

[31] This is only *suggested* by the authorities, eg in *Oye* [2014] 1 All ER 902.

[32] Criminal Appeal Act 1968, s 6. This approach has been championed by some leading authorities: *Blackstone's Criminal Practice* (n 29) [A3.23]: 'In principle, where the accused's lack of responsibility is caused by insanity rather than any other factor, it would seem logical that the defence should be confined to, and classified as, insanity irrespective of whether the lack of responsibility takes the form of no mens rea/automatism or a belief in a justifying defence'. See too Law Commission, *Insanity and Automatism* (n 4) [2.18]-[2.29]. The other way to avoid pleading insanity, of course, is simply for D to plead guilty, as occurred in *Sullivan*.

[33] This is criticised by Simester et al, *Simester and Sullivan's Criminal Law* (4th edn, Hart 2010) 714 fn 83, cited in Law Commission, *Insanity and Automatism* (n 4) [2.10].

[34] The prosecution must still prove the actus reus to avoid an outright acquittal: *AG's Reference (No 3 of 1998)* [2000] QB 401.

[35] *AG's Reference (No 2 of 1992)* [1994] QB 91; *Coley* [2013] EWCA Crim 223 [22].

[36] Simester et al, *Simester and Sullivan's Criminal Law* (6th edn, Hart 2016) 118-119 criticise this restriction for this reason, preferring the more liberal line of authorities represented by *Charlson* [1955] 1 All ER 859 and *Quick* [1973] QB 910. Another possible restriction to automatism is that it is unavailable when (recklessly) self-induced: *Quick*; *Bailey* [1983] 1 WLR 760; *Hardie* [1985] 1 WLR 64. Only 'possible' for two reasons. First, it's not clear whether automatism would be permitted for crimes of *intent* if *recklessly* self-induced. The better interpretation is probably that the self-inducement must come with the grade of 'mens rea' required for the relevant offence. (See, eg *Blackstone's Criminal Practice* (n 29) [A3.14].) Second, however, even if the more liberal position is right, recklessly self-inducing automatism is hardly

In both cases, the incapacity doctrines turn what would be straightforward acquittals via denials of offending into much riskier trials that might lead to the special verdict or even criminal liability. If we think that incapacities exculpate then in these cases we ought to prefer the objector's proposal. That proposal, unlike the status quo, at least guarantees an acquittal in the absence of mens rea.

Three responses. First, and most importantly, only some incapacity doctrines have this inculpatory effect. Infancy exculpates no matter whether the defendant did or did not possess the requisite mens rea. The same applies (albeit for different reasons) regarding unfitness to plead. At best, then, this objection is equivocal as to the *net* generosity of the incapacity doctrines.

Second, we can agree that any inculpatory effect is problematic while insisting that the solution is reform rather than abolition. The reason why insanity has its quasi-inculpatory structure is because of a perceived gap in public protection should insane defendants receive complete acquittals.[37] While this gap is perhaps exaggerated,[38] the best response would be to fill it—perhaps with stronger civil powers—rather than ignoring it. Unlike the objector's proposal, this would address those policy concerns while protecting the minority of defendants who rely on the incapacity doctrines despite having fulfilled all the elements of an offence.

Third, the discussion above highlights a yet more subtle category of incapacity rule. The objector pointed out that insanity, and possibly automatism, operate like intoxication: they allow the prosecution to impute (a substitute for) mens rea where no mens rea is proven. These rules, along with prior fault rules, are *imputation* doctrines. They allow the prosecution to impute elements they have not proved in the usual way. The law could endorse a much wider range of imputation doctrines. It already strips defendants of the benefit of the defence of duress if the defendant had certain prior associations with their duressors.[39] It denies the automatism defence where it was recklessly self-induced.[40] It could remove the need to prove any number of elements or sub-elements in any number of cases. The fact that the law resorts to imputation doctrines only very rarely implies a preference against ignoring relevant particularities in each

---

equivalent to recklessly committing the full offence. Once again, the automatism doctrine serves, if anything, to restrict the scope of denials of offending.

[37] Law Commission, *Insanity and Automatism* (n 4) [2.25]-[2.29].

[38] See my criticism of *C (Sean Peter)* [2001] EWCA Crim 1251 in ch 2 §3.2.

[39] *Z/Hasan* [2005] 2 AC 467.

[40] *Hardie* and *Bailey* (n 36).

case, a kind of *anti-imputation presumption*. One reason for this presumption might be to account for defendants' incapacities. The more imputations the law adds the smaller the scope is for genuine incapacities to be given their full weight in the final assessment of defendants' conduct. This anti-imputation presumption constitutes a weaker cousin of counterfactual relativisations. It is too weak, and the identification of the grounds for such a presumption too speculative, to promote to the full ranks of my taxonomy of incapacity rules. But the objections above help us to see how the law accounts for incapacities in subtler ways still than those identified in my taxonomy.

2 Proxies

Some incapacities make offenders less culpable. The criminal law accounts for these incapacities with its incapacity rules. The incapacity rules do so by identifying *proxies* for the relevant incapacities: young age, recognised mental illnesses, etc.[41] We cannot move from the claim that certain incapacities exculpate to the conclusion that the incapacity rules are justified without an additional argument to justify the use of these proxies. Gideon Yaffe has recently argued that such additional arguments fail.[42] He focuses on *age* proxies, but his objection could extend to the other incapacity rules.

Here's how Yaffe formulates the additional argument we need to justify proxy-based incapacity rules (for age):

> *The proxy for culpability argument* (PC)
>
> (PC1) Kids lack feature F[43]
>
> (PC2) Agents are fully responsible for their wrongful conduct only if they have feature F […]
>
> (PC3) Agents should be given a break for their wrongful conduct if they are not fully responsible for it

[41] Insanity and diminished responsibility do not use recognised mental illnesses as *pure* proxies; they also require evidence of certain underlying incapacities. But we don't fully understand how best to formulate the relevant incapacities, nor how exactly they bear on culpability. That means we can interpret these rules' reference to incapacities as *oblique* proxies for *the correct conception of the underlying relevant incapacities*.

[42] Gideon Yaffe, *The Age of Culpability* (Oxford 2018) introduction, ch 1.

[43] Yaffe suggests normative competence and susceptibility to peer influence as relevant attributes, ibid 22-23.

(Conclusion) ∴ Kids ought to be given a break for their wrongful conduct[44]

(I'll return to what he means by 'give kids a break'. For now, call it 'age-based leniency').

Yaffe's objection to PC focuses on the fact that not all 'kids' in (PC1) fall within the class of 'agents' in (PC2) and (PC3). (PC1) is only a rough generalisation. There is no bundle of natural attributes that plausibly bear on responsibility that *all* kids lack.[45] Age is only a proxy for those-who-lack-F. But, in that case, why not simply give *those-who-lack-F* a break? Why care about kids, specifically? As Yaffe puts it, an advocate of PC 'must supplement [PC] substantially in order to reach the further conclusion that we should *adopt a social policy* of giving kids a break.'[46] It is not PC per se, but rather this supplementary argument that Yaffe criticises. Reconstructed, the relevant supplementary argument is:

Best proxy argument (BP)

(BP1) It is justified to target F via a proxy
(BP2) Ceteris paribus, it is only justified to use the *best* proxy[47]
(BP3) Age is the best proxy to target F
(Conclusion) ∴ Ceteris paribus, it is justified to target F via age as a proxy

Yaffe can grant (PB1) arguendo: proxies are often justified.[48] But it doesn't follow that some specific proxy is justified. Yaffe sets a high standard: only the *best* proxy is justified (BP2). But (BP3) is clearly false: proxies other than age (alone) could better target F. Ceteris paribus, a proxy is better if more accurate. They are more accurate if less under- and over-inclusive. The age proxy would be less under-inclusive if it included some immature adults. It would be less

[44] ibid 24.

[45] Nor can we interpret F *analytically* without rendering the premises (respectively) trivial and false. That is, if F were 'the feature of being 18 years old,' the (PC1) would read 'Kids are kids' (trivial), and (PC2) 'Agents are less responsible only if they are kids' (false): ibid 24-26.

[46] ibid 26 (emphasis in original).

[47] ibid 26-29. Yaffe cashes out 'best' as 'the overall value generated by the [proxy policy] …is greater than that of equally implementable policies', where the 'overall value' includes the value of true positives and negatives minus the disvalue of false positives and false negatives.

[48] (Most) laws are rules and rules are proxies. While this is obvious for speed limits (target: bad driving), it's also true of paradigmatic *mala in se* crimes like homicide (target: impermissible killings). The latter remain proxies as there are permissible killings (necessity, mercy, etc) which homicide laws nonetheless prohibit. We can't reformulate these offences to capture only *killings-we-don't-want* without making them offensively vague. There is no rule-of-law compliant solution to over- and under-inclusivity at the margins. Indeed, prohibiting *killings-we-don't-want* would likely fare badly even at *preventing/punishing killings-we-don't-want*, as a result of providing insufficient guidance for officials. In general, see Frederick Schauer, *Playing by the Rules* (Oxford 1991). (Though, as Schauer points out at 11, not *all* laws are rules; some invoke open-ended standards).

over-inclusive if it excluded some mature kids.[49] The law doesn't do this. We wouldn't want it to. And this, Yaffe claims, means we would not use the best proxy, and thus that PC (supplemented with BP) does not capture our rationale for giving kids a break.

Yaffe doesn't dwell on under-inclusivity. But it is much less of a problem than he suggests. The law *does* use additional proxies. It doesn't just give 'kids' a break. It gives breaks to the insane, to those with recognised mental illnesses, to those who lack various incapacities. That was the lesson of chapter 3. Yaffe is aware of these rules. But he takes them to favour *his* argument. He claims that

> we do not need to add a policy of leniency towards kids to the law in order to give a break to… [less blameworthy] people; *they are already given a break under the criminal law*. A person who was not vividly aware of the risks of harm that his act caused will often be characterized as negligent, rather than reckless, and subject to lesser penalties. A person who acted impulsively, or without sober reflection on the pros and cons of conduct, will sometimes have a less objectionable grade of mens rea… The law already recognizes a wide range of excuses, tailored to approximate… moral differences in blameworthiness.[50]

This would be true if the law's generally applicable excuses (and gradations between offences, etc) adequately accounted for age-derived exculpatory factors. But that is not true. First, in some contexts, we are entitled to expect less from people with certain attributes, including young age. By default, however, the law's standards (like reasonableness) do not account for these attributes. To account for them, the law must relativise its standards to account for specific attributes. The generally applicable excuses (etc) do not accommodate age-derived exculpatory factors without specifically age-based provisions.[51] Second, the insanity defence is the law's ultimate safety valve to avoid convicting incapably non-culpable defendants. Defendants only count as insane if their incapacity derives from a specific aetiology: a

[49] Yaffe, *The Age of Culpability* (n 42) 26-30, and esp. 31-32.

[50] ibid 7.

[51] See ch 2. Doesn't Yaffe deny that there are age-derived exculpatory factors? No. His claim is just that such factors are already accounted for via other doctrines. Children may be less culpable as less reckless as they don't foresee risks. But we don't need an age-based explanation for this exculpation: a general explanation in terms of foresight will do. As I said in chapter 2, this is misleading if the rationale for omitting a standard was to precisely to account for age, that is, for counterfactual relativisations. But it is simply false when it comes to relativisations proper. Take duress. We do not expect a child to demonstrate the reasonable firmness of an adult. But, unless we specifically relativise 'reasonable' to age, that is precisely what the law would demand.

recognised mental illness. An adult with the mental age of a 7-year-old child has the necessary aetiology, counts as legally insane, and is thus given a break. But an *actual* 7-year-old child's equivalent incapacity does *not* constitute a recognised mental illness. That child thus lacks the necessary aetiology to count as insane and is not given a break. This child is deprived of a defence available to an otherwise identical adult.[52] To fill this lacuna, the law must accept age as an *additional* qualifying aetiology. The lesson from both relativisations and insanity are that the law must use age in addition to the generally applicable excuses (etc). Once we see this, it's easy to see that age need not be the best proxy for F standing alone. Age-based rules need only *contribute* to the kaleidoscope of incapacity rules that *collectively* constitute the best proxy for F. Under-inclusivity is not a problem for BP once we modify (BP3) to (BP3*): age *contributes to* the best proxy to target F.

But Yaffe is right when it comes to over-inclusivity. Some developmentally mature children will possess F. We could exclude (some of) these children from our proxy. That would make our proxy more accurate. Ceteris paribus, a proxy is better if more accurate. Thus (BP3*) needs further modification to (BP3*+): age-*and-maturity* contribute to the best proxy to target F. BP should conclude that 'it is justified to target F via age-*and-maturity* as a proxy.' PC would then exclude mature kids from the proxy it uses to justify giving kids a break. Ultimately, Yaffe's objection is to this exclusion. Reconstructed:

Yaffe's objection (YO)

1) Some kids possess F
2) Following PC, we would not give a break to kids who possess F
3) We would (and ought) to give *all* kids a break[53]
4) ∴ We would (and ought) not to follow PC

'Following' PC in (2) includes adopting (BP3*+). That entails excluding mature kids from our proxy. That, in turn, entails not giving all kids a break, contra (3). The problem for Yaffe is that PC advocates explicitly reject (3). The entire point of proxy arguments is that the target

[52] This is the truth in the 'kids will be kids' argument Yaffe rejects in Yaffe, *The Age of Culpability* (n 42) ch 2. The developmental normality of kids' incapacities *deprives* them of a defence otherwise available. The law doesn't give them a *new* break to account for developmental normality. Rather, it reinstates an *existing* break that was *denied* because of developmental normality. (Some children may still qualify as legally insane (eg if they have certain psychiatric disorders); the point is that most young children are functionally 'insane' but lack the necessary aetiology).

[53] ibid 30, 33-34. As Yaffe summarises at 39, 'The policy's appropriateness is not empirically dependent'.

attribute F, not the proxy (here, age), is what ultimately matters.[54] To vindicate (3), then, Yaffe tries to demonstrate that, contra our expressed views, PC advocates *would* give all kids a break, *even if not doing so would make for a better proxy*. If so, it follows that we do not really endorse BP (as modified), nor in turn PC.

(I focus on the argument that we *would* give all kids a break and parenthesise the claim that we *ought* to do so, as Yaffe (by his own admission) simply assumes the latter).[55]

Yaffe's argument for (3) runs as follows. Plausibly, a proxy for F sensitive to *gender* or *height* may be more accurate than age alone. However, we wouldn't endorse a proxy for F sensitive to gender or height. Thus, we do not endorse the best possible proxy. Thus, our rationale for giving kids a break cannot be that age contributes to the best possible proxy.[56]

The PC advocate has two responses. First, we might object to using gender or height proxies for reasons unrelated to their accuracy qua proxies: reasons of equality, discrimination, etc. The *best* proxy is not reducible to the *most accurate* proxy. But Yaffe rightly points out that this only gets us so far. Reasons of equality, discrimination, etc do not set absolute constraints. The value of predictive accuracy *might* outweigh the disvalue of using such proxies. Second, then, we must bite the bullet. It *might* be justified to use gender or height proxies if the gain in

---

[54] More precisely, PC advocates would not endorse policies that apply to *all* tokens of an imperfect, empirically dependent proxy where a better proxy is available. Yaffe calls this the 'problem of empirical dependency'. According to (3), an adequate explanation for giving kids a break must output the conclusion that all kids are owed breaks. But we could explain how age is a *perfect* proxy consistently with (3). That, indeed, is Yaffe's view. Roughly, he argues that age is a proxy for *a suitable amount of elapsed time in which parents can influence the views of future citizens*, which is in turn a proxy for ensuring the ultimate values of *political equality and self-government* (ibid ch 6). My concern, however, is only with his negative thesis.

[55] It is a practice Yaffe is '*certain has a good rationale*' (emphasis in original) (ibid 3), that 'cannot be denied' (33-34; 183). But consider some futuristic cases.

> *Emulated*: D is emulated at a high speed. D is 5 in earth years but developmentally 50.
>
> *Relativity*: Baby D takes a round-trip in space close to the speed of light. Due to relativity, D returns as a developmental infant, but many earth-years later.
>
> *Modified*: D, genetically modified, is indistinguishable from an adult by age 10.

In these cases, it seems like D's age (spacetime location) is morally irrelevant to whether we should give her a break. What matters is her development. That conclusion transfers to ordinary cases, like

> *Mature*: D is remarkably mature for her age.

It seems plausible to me that D's maturity, not her age, is what matters, and thus that we should reject Yaffe's assumption. (*Emulated* is based on Robin Hanson, *The Age of Em* (Oxford 2016) ch 6; *Relativity* on Tyler Cowen, *Stubborn Attachments* (Stripe 2018) 68.)

[56] Yaffe, *The Age of Culpability* (n 42) 31-33. As Yaffe glosses the assumption, age is 'intuitively ethically sticky…when it comes to criminal responsibility' (33).

accuracy is sufficiently valuable to defeat the disvalue of using those categories. This bullet Yaffe refuses to bite. Nor does he acknowledge that we might. Far from taking seriously gender-based proxies for F, he thinks the idea should be 'laughed out of contention'.[57] The PC advocate might join him, but only via our first response: perhaps the discriminatory disvalue of gender proxies is so great as to outweigh any value gained from improved accuracy. But Yaffe doesn't accept this response. He thinks that gender-based discrimination may plausibly improve accuracy sufficiently to outweigh its discriminatory disvalue *and yet still be impermissible*. He presses us to agree, and thus to renounce BP, and in turn PC.

To clarify the dialectic, both Yaffe and PC advocates accept:

(P) It is unjustified to use certain proxies to target F

PC advocates explain *which* proxies are unjustified by appeal to:

(Q) It is unjustified to use a proxy if and only if the disvalue of its use defeats the value of any gain in accuracy

But Yaffe claims that we implicitly endorse a conflicting premise:

(R) It is unjustified to use some proxies where the disvalue of their use *does not* defeat the value of any gain in accuracy

The difficulty with Yaffe's argument is that PC advocates can offer an expansive account of the relevant disvalues.[58] If we have a strong intuition (robust under reflective equilibrium, etc) that some proxy is impermissible, we can always cash this out in terms of its disvalue.[59] Then we can happily endorse that intuition consistently with (Q). To this Yaffe has two responses.

---

[57] This is most obvious if positive discrimination justifies the choice of *less* accurate proxies to rectify past discrimination. It's also clear where there the discrimination is justified by some proportionate end. Yaffe gives the example of disproportionately hiring women applicants for airport staff who pat down travellers, as travellers may only be comfortable with same-gender staff. ibid 36. Similarly, the Equality Act 2010's protections of certain characteristics allow many exceptions for proportionate ends. (Part 14, and various Schedules).

[58] Note that 'defeats' is broad, including exclusion or other second-order reasons. Yaffe, *The Age of Culpability* (n 42) 38-39. I follow Yaffe in cashing out the idea in terms of value and disvalue, but we could equally phrase it in terms of reasons for and against.

[59] There is an apparently simpler problem with Yaffe's argument: it's analytic that if the value of A is undefeated by the disvalue of B, then A is permissible. (Q) is coherent, (R) incoherent. (There are complications of commensurability, etc, but I find this picture of value theory plausible, and the point works mutatis mutandis even accounting for those complications). But I think Yaffe's point is not that we *should* endorse (R). Rather, it is that we PC advocates really do implicitly endorse the incoherent (R) and renounce the coherent (Q). At that juncture we could just thank Yaffe for pointing out our mistake, and happily (re)endorse (Q), ie bite the bullet. But Yaffe's claim is that we would not do this. Instead, this incoherence implies a mistake upstream in the argument. That mistake is trying to justify age *as a*

First, Yaffe uses intuition pumps to suggest that some unjustified proxies cannot be explained in this way. He accepts that some proxies are ruled out by the disvalue of historical discrimination (eg race).[60] But he asserts that this cannot explain our aversion to gender proxies, as 'it would be false to say that women have been more oppressed by the criminal law overall than men have; the reverse seems much more unlikely.'[61] Likewise, he suggests that

> For all we know, some childhood deprivations might increase the rate at which people acquire property F…[as] we know that tough circumstances speed other kinds of learning… If that turned out to be true, then a policy that denied a break to, say 16- and 17-year-olds, provided they grew up in poverty…might offer a better mix of true and false positives.

Yaffe continues that we PC advocates wouldn't exclude from our proxies more-capable women and deprived children, and thus must be committed to (R) rather than (Q). But Yaffe's examples are highly implausible. He offers no evidence for either claim. As for gender, yes: men are convicted at higher rates than women. But what of the long history of differential and insufficient criminal *protections* for women, not least a long history of chattel status?[62] Couldn't this lamentable history explain the enormous disvalue of differential treatment today?[63] The

---

proxy. The better view, says Yaffe, is that age plays a different role in the argument. (ibid 33). My response in the text therefore aims to vindicate (Q) by showing that it is not just coherent but also attractive.

[60] Yaffe limits his examples of disvalue to historical discrimination, and his examples of potentially justifying values as the fulfilment of 'essential government activity' (ibid 36-37). It's not clear whether Yaffe's examples were intended to be exhaustive, but I can see no reason to accept that view. All types of values matter.

[61] ibid 38.

[62] Yaffe's neglect of the criminal law's *protections*, as opposed to its *sanctions*, leads him to suggest that a history of racial discrimination might make it 'better that one hundred guilty Black people go free than that one innocent Black person be convicted; or perhaps the right ratio is 1000:1, even though 10:1 does just fine for the population as a whole.' He neglects the likely effect of this policy: drastically and differentially failing to protect black *victims*. According to US uniform crime reporting (UCR) program data, in 2016 of 2870 black murder victims (in single victim/single offender cases), 2570 (90%) had black perpetrators. (The intra-white rate was 82%). In 2017 the rate was 88% for blacks, 80% for whites: see FBI, *Crime in the United States 2016*, Expanded Homicide Data Table 3: < https://web.archive.org/web/20181207191656/https://ucr.fbi.gov/crime-in-the-u.s/2016/crime-in-the-u.s.-2016/tables/expanded-homicide-data-table-3.xls> archived 7 December 2018; FBI, *Crime in the United States 2017*, Expanded Homicide Data Table 6: < https://web.archive.org/web/20181029184714/https://ucr.fbi.gov/crime-in-the-u.s/2017/crime-in-the-u.s.-2017/tables/expanded-homicide-data-table-6.xls> archived 29 October 2018. The self-report National Crime Victimization Survey (NCVS) consistently finds that a majority of violent victimisation is intraracial. Rachael Morgan, 'Race and Hispanic Origin of Victims and Offenders, 2012-15' (2017) NCJ 250747: <http://www.bjs.gov/index.cfm?ty=pbdetail&iid=6106> accessed 15 December 2018.

[63] Where to start? A premise of early medieval English law was that 'public law gives a woman no rights'. Frederick Pollock and FW Maitland, *The History of English Law Before the Time of Edward I* (2nd edn, Cambridge 1898/Liberty Fund 2010) bk II ch II §11 (465/508). Later medieval byelaws frequently

same goes for poverty. It is overwhelmingly more likely that childhood deprivation would have precisely the *opposite* effect that Yaffe suggests: to delay acquisition of the relevant capacities.[64] And, even if he was right that excluding the impoverished would improve accuracy, the value of that gain must be amply defeated by the disvalue of compounding injustice through poverty-based exclusions. These intuition pumps provide no support for Yaffe's argument. We can easily explain our aversion to these proxies by reference to their (at best) low gains in accuracy and their high accuracy-independent disvalue. They give us no reason to abandon (Q) for (R).

Second, we might still worry that this response is too easy. If PC advocates rely on an expansive account of the relevant disvalues to rule out certain proxies, on what basis does it rule *in* age? Is there not a long history of age-based discrimination? Isn't this disvaluable enough to rule out BP, and in turn PC? To answer this, the PC advocate will consider the value of any accuracy increase and weigh that against the disvalue of discrimination. Yaffe's argument only succeeds if this standard balancing exercise doesn't offer a plausible, non-ad-hoc explanation for including age but excluding gender, height, race, etc in its proxy. But the standard balancing exercise offers a perfectly plausible answer to why we include age but not gender (etc). The argument is simple: the gain in accuracy by using age within our proxy is *very high*, and in turn *very* valuable. As we saw above, without including age as a proxy the law would fail to account for a large category of individuals lacking F: children. By contrast, there is no such large gain to be had by using gender, height, race, etc as a proxy. Given the enormous value of using age as a proxy, we need not explain away the disvalue of discrimination. It is simply outweighed. But, as it happens, the law's historical treatment of children has also been largely beneficent. Compared to women, ethnic minorities, etc, the case against differential treatment for children is much weaker. The standard PC explanation for including age as a proxy is entirely plausible: it achieves great value with little disvalue. Again, we need not abandon (Q) for (R).

---

outlawed (and 'vivid[ly]' punished) 'whores', a status (not an activity) attaching only (and quite indiscriminately) to women: Ruth Mazo Karras, *Common Women: Prostitution and Sexuality in Medieval England* (Oxford 1996) ch 1. Within recent memory the English law of rape required not just non-consent but proof of force, fear, or fraud (until *Olugboja* [1982] QB 320) and exempted husbands from liability for raping their wives (until *R* [1992] 1 AC 599). For a global survey of modern gender discrimination, see Martha Nussbaum, *Women and Human Development: The Capabilities Approach* (Cambridge 2001) 1-4 and references therein.

[64] It seems likely that childhood poverty results in increased impulsivity/temporal discounting, itself correlated with crime. The results of the famous marshmallow test (childhood delayed gratification predicts better life outcomes), which Yaffe mentions 55-56, are likely best explained by poverty: Tyler Watts, Greg Duncan, and Haonan Quan, 'Revisiting the Marshmallow Test: A Conceptual Replication Investigating Links Between Early Delay of Gratification and Later Outcomes' (2018) 29(7) Psych Sci 1159. For a popular treatment of the hypothesis, see Sendhil Mullainathan and Eldar Shafir, *Scarcity* (Times Books 2013).

Yaffe has a final objection to my claim that using age as a proxy is valuable for giving children a break. That's because he does not define 'kids' as 'children'. Rather, he defines 'kids' as 'under-18s'. Similarly, by 'breaks' he does not have in mind only, or primarily, the infancy defence. Rather, he means *any* form of lenient treatment, including milder processes and sentences.[65] Using those definitions, it *does* seem that ruling in age and ruling out gender seems quite arbitrary. It's hard to imagine that we gain much by way of accuracy in targeting F by offering sentence reductions to 17-year-olds rather than 18-year-olds, and it does seem invidious, or at least arbitrary, to distinguish those two classes. Given that we do offer 'breaks' specifically to 'kids' (so defined), does this not rehabilitate Yaffe's objection?

Not really. For one thing, there is simply no one 'policy' nor set of policies that offer general lenience to the class of under-18s (or any other discrete age bracket). English sentencing law contains a bewildering array of age distinctions for various offence and sentence types, often using three categories of 10-17, 18-20, and 21+.[66] Different categories are used for substantive criminal law: under 10 for infancy, 10-13 for the (abolished) doli incapax defence.[67] Meanwhile, both sentencing and the substantive law (especially relativisations) permit scalar accounting for age. What counts as 'reasonable firmness' may differ between 13- and 16-year olds, thereby affecting mitigation or the availability of the defence of duress. In short, the law's policies are far more fine-grained than Yaffe's definitions imply. They do account for different age brackets. But why does the law finely attune different policies to different age brackets? Here the problem with Yaffe's definitions become obvious: the law uses different policies better to track distinctions in culpability. Under 10s are almost invariably incapably non-culpable. By contrast, there is greater variance among 17-year olds, whose average capacity is much higher. It makes perfect sense, then, to offer a categorical exemption to under 10s (via infancy) and cautious case-specific leniency to 17-year olds (via sentencing distinctions). This is perfectly explained by PC. Exempting under-10s from all liability is very accurate; exempting under-17s would be massively over-inclusive. We do not have 'a policy' of 'giving kids a break'. We have various policies giving various breaks to various classes of kids, all finely attuned to culpability differences between those classes. Above I said that age *contributes* to the broader proxy for targeting F: the incapacity rules. In fact, the truth is more nuanced still. *Different* ages contribute

[65] Yaffe, *The Age of Culpability* (n 42) 1-2.

[66] Julian Roberts, Nicola Padfield, and Lyndon Harris (eds), *Current Sentencing Practice* (Sweet & Maxwell 2018) SR1-150. The law currently specifies that the aim of youth justice is to prevent reoffending, and that criminal courts must have regard for children's' welfare: Crime and Disorder Act 1998 s37, Children and Young Persons Act 1933 s44.

[67] Ch 2 §2.2.

in different ways to that proxy. And this justification, suitably understood, is very well explained by PC.

My reason for avoiding certain proxies is adequately explained by (Q). If there were some better proxy that did not entail giving all kids a break, I would accept it. I reject (3). I see no reason to abandon BP or PC. We can justify the incapacity rules as proxies for lowered culpability.

3 Aetiologies

We *can* justify the incapacity rules as proxies for lowered culpability. But *are* the law's incapacity rules so justified? This claim requires us to distinguish between the underlying determinants of defendants' lower culpability—what Yaffe called feature F—and the proxies we use to identify those determinants. Given the focus of our inquiry, we're concerned with only a subset of those determinants: *incapacity-based* lowered culpability. Above we focused on age as a proxy for (incapacity-based) lower culpability. But, I claimed, young age was but one proxy among others. The most prominent additional proxy is mental disorders. They're relevant to insanity, diminished responsibility, and fitness to plead. But those incapacity rules don't use mental disorders in the way that infancy uses young age. Proof of young age *substitutes for* proof of feature F. It's a proxy for incapacity (and thus lower culpability). By contrast, proof of mental disorder is required *in addition* to direct proof of incapacity (and thus lower culpability). These incapacity rules use mental disorders as an additional *aetiology requirement*. These aetiology requirements present a justificatory challenge for the incapacity rules.

Consider the now-familiar formulation of the insanity defence from the M'Naghten rules:

> [T]o establish a defence on the ground of insanity, it must be clearly proved that, at the time of the committing of the act, the party accused was labouring under such a defect of reason, from disease of the mind, as not to know the nature and quality of the act he was doing, or, if he did know it, that he did not know he was doing what was wrong.[68]

Separated out, one variant of the defence is permitted if it is proved that:

1A) D did not know he was doing what was wrong [caused by]

[68] *M'Naghten's case* (1843) 10 Cl & Fin 200, 210.

1B) D's defect of reason [caused by]

2) D's disease of the mind

We saw in chapter 2 that the M'Naghten rules are best interpreted as referring to an *incapacity* to understand wrongfulness, not merely ignorance of wrongfulness.[69] That interpretation is plausibly justified by the language of a defect of reason. A better formulation of the rules could therefore combine (1A) and (1B) to make this explicit.[70] Then we could distinguish more cleanly between the underlying culpability-lowering incapacity and the aetiology requirement:

1) D was incapable of knowing that his act was wrong [caused by]

2) D's disease of the mind

The aetiology requirement picked out by (2), a disease of the mind, is a plausible *proxy* for inferring that D might be incapable of knowing that his act was wrong, the incapacity picked out by (1). But the law of insanity doesn't settle for proof of (2) in lieu of proof of (1). It demands proof of that aetiology *in addition* to proof of (1). Other incapacity rules work in a similar way.[71]

Why is this a problem? Consider infancy again. Children are morally exculpated on the basis that they cannot—in general—distinguish right from wrong. Likewise, we might assume, for insane people. That means that step (1) identifies the basis for exculpation, feature F. But once we have direct proof of feature F there's no longer any need for a proxy for it. Demanding proof of (2) then seems quite redundant. Worse, in fact. Imposing an aetiology requirement seems to *deny* a break to provenly morally exculpated defendants.

The puzzling nature of the aetiology requirements is emphasised by the courts' staunch rejection of any *additional* aetiology requirements to the insanity defence. Take *Sullivan*. Sullivan

[69] Ch 2 §2.1

[70] A defendant cannot benefit from the insanity defence if they suffered a disorder that had no bearing on their conduct. The accused must have been labouring under a defect of reason *so as not to know* the nature of their conduct or that it was wrong. That causal nexus is crucial. The Butler Committee report raised the suggestion of abandoning that nexus, but the proposal found few supporters.

[71] Diminished responsibility is available only if the defendant prove not only (say) a substantially impaired ability to understand the nature of their conduct (their exculpatory incapacity), but also that this incapacity stemmed from an abnormality of mental functioning arising from a recognised medical condition (the aetiology requirement): Homicide Act 1957 s2(1) (as amended by the Coroners and Justice Act 2009 s52). In duress, medical evidence that one was vulnerable to pressure is inadmissible unless that vulnerability stemmed from a mental illness which, additionally, generally produces such effects. This is not just an aetiology requirement, but a *common* aetiology requirement. *Bowen* [1997] 1 WLR 372, 379 (principle 5).

kicked and injured an elderly old man during an epileptic fit. He wanted to plead (non-insane) automatism, and thus receive a straight acquittal. But his incapacity derived from an internal cause, and thus could only count as insanity. In response, Sullivan argued that epilepsy doesn't constitute a disease of the mind (and therefore isn't insanity) in ordinary parlance. Insanity means something more like a permanent psychiatric disorder. If Sullivan's argument was correct, however, the M'Naghten rules would require an additional element, something like

3) [D's disease of the mind was caused by] a permanent psychiatric condition

Lord Diplock rejected his argument. Why? Because

> it matters not whether the aetiology of the [disease of the mind] is organic, as in epilepsy, or functional, or whether the impairment itself is permanent or is transient and intermittent, provided that it subsisted at the time of the commission of the act.[72]

This seems harsh to Sullivan. Adding (3) would have opened the path to a straightforward acquittal. But for other defendants, the alternative to the insanity defence would just be criminal liability.[73] Adding (3) might result in criminal liability for someone who acted in the grip of transient psychosis. Lord Diplock's refusal to add an additional aetiology requirement like (3) shields such defendants from liability. This shielding is appropriate because the basis for their lesser culpability is their incapacity, *regardless of its aetiology*. The Law Commission has made the same point in defence of their proposed reforms:

> the reason for a defence based on…non-responsibility would be the lack of capacity. The background causes of that lack of capacity are not legally relevant...[74]

The point made by both Lord Diplock and the Law Commission is that the aetiology of a defendant's mental disorder is morally irrelevant. Requiring proof of additional aetiologies, such as (3), would be a mistake. But, in that case, and in the very words of the Law Commission, isn't the aetiology of a defendant's *lack of capacity* itself morally irrelevant? And,

---

[72] *Sullivan* [1984] AC 156, 172.

[73] Eg if D intentionally harmed V while sleepwalking in the sleep-induced belief that V was a vampire and that such harm was legally mandated.

[74] Law Commission, *Insanity and Automatism* (n 4) [2.18] (regarding fetal alcohol syndrome, citing Sanford Kadish, 'Excusing Crime' (1987) 75 Calif L Rev 257). The Law Commission add an important caveat, to which I'll return below.

if so, why does the law demand proof that D's incapacity, step (1), derived from *any* specific aetiology, including (2)? What justifies these aetiology requirements?

I'll consider three possible justifications: (i) that certain aetiologies *morally transform* incapacities, *making* them exculpatory; (ii) that non-qualifying aetiologies are disqualified in virtue of *prior fault*, and (iii) that the aetiology requirements improve the law's proxy for culpability by *reducing error*.

3.1 Moral transformation

The moral transformation reply aims to dissolve the puzzle of the aetiology requirements. I claimed that an incapacity to understand wrongfulness, step (1), was itself a basis of moral exculpation, that is, (a subset of) what Yaffe called feature F. But this might have been premature. Perhaps the wrongness limb does not, by itself, amount to a ground of moral exculpation in the case of insanity. Perhaps (1) is morally exculpatory *only if* it derives from certain aetiologies, aetiologies like (2). Then, far from being mere surplusage, that aetiology is *constitutive* of an incapacity being morally exculpatory in the first place.[75] The aetiology *morally transforms* an otherwise non-exculpatory incapacity into an exculpatory incapacity. Thus: aetiology requirements are necessary for incapacity rules to pick out genuinely morally exculpatory incapacities. Puzzle dissolved: (2) is necessary because morally transformative.[76]

In a sense, I've already endorsed a version of this argument. I took two elements of the M'Naghten rules:

1A) D did not know he was doing what was wrong [caused by]
1B) D's defect of reason

and reduced them to

1) D was incapable of knowing that his act was wrong

[75] This specifies a necessary but not a sufficient condition for an incapacity to be exculpatory. I consider what makes incapacities exculpatory in general in ch 5.

[76] This explains the intuition that it would be insufficient for a defendant, upon being accused of some crime, to answer: 'But I lacked a relevant capacity.' We expect another part to the answer which explains that incapacity, such as 'I was only 9 years old' or 'I was suffering from a severe mental disorder'. But while this intuition implies that some further aetiological explanation ought to be offered, it doesn't imply that there is a fixed list of qualifying aetiologies. It is the use of this limited list of qualifying aetiologies which stands in need of justification.

I did this to achieve an explicitly incapacity-based formulation, given that that is how the M'Naghten rules are typically understood. But, leaving that analysis to one side, (1B) looks like an aetiology requirement. The law will only accept moral ignorance as exculpatory if it derives from a specific aetiology: a defect of reason. Does (1B) therefore raise the same justificatory challenge as other aetiology requirements, including (2), the disease of the mind?

I don't think so. Why not? Because mere ignorance of moral wrongfulness is not ipso facto exculpatory. Ignorance is exculpatory only because it arises from a defect of reason.[77] Why? Because a defect of reason renders defendants not *merely* ignorant, but also *incapable* of knowing that their conduct is wrong.[78] It is this incapacity that is exculpatory. The defect of reason morally transforms mere ignorance into an exculpatory incapacity.

Can an analogous argument work for (2)? Might (2), a disease of the mind, be required for (1), the incapacity to distinguish right from wrong, to count as a ground of exculpation? Not so fast. The cases are importantly disanalogous. (1A) alone refers to mere ignorance. It doesn't specify *any* kind of incapacity. It's no puzzle for my account that mere ignorance needs supplementation to exculpate.[79] It is a puzzle, however, why an exculpatory incapacity needs supplementation, and (1) looks like it picks out just such an exculpatory incapacity. That's why (2) raises a justificatory puzzle but (1B) does not. The moral transformation response only dissolves the puzzle raised by (2) if, contrary to appearances, (1) does *not* pick out an incapacity that is ipso facto exculpatory.

---

[77] Or due to some other exculpatory factor beyond the mere fact of ignorance.

[78] True: the M'Naghten rules don't actually say this. There could be defendants who are morally ignorant due to a defect of reason and yet who remain *capable* of understanding the wrongfulness of their conduct. Under a literal reading of the M'Naghten rules (and the model jury directions based on them) such defendants may yet be entitled to the insanity defence. Two responses. First, as I explained in chapter 3, most glosses do use explicitly incapacity-based formulations. This brings insanity in line with other analogous incapacity rules. Diminished responsibility refers to an 'impaired *ability*' to form a rational judgment, while unfitness to plead asks whether D was '*capable* of…understanding the charges'. (Insanity is distinguished from intoxication on this score: *Cole* [1993] Crim LR 300). Second, if there is a gap, consider my reduction a reforming suggestion. Such capable defendants *ought not* to be entitled to plead insanity. The appeal of this suggestion will depend on how we understand what it means to be relevantly incapable, which I discuss below.

[79] It might raise a puzzle for other moral philosophers, especially if that ignorance meant that D lacked mens rea. See, eg Gideon Rosen, 'Culpability and Ignorance' (2003) 103 PAS 61; Gideon Rosen, 'Kleinbart the Oblivious and Other Tales of Ignorance and Responsibility' (2008) 105 Journal of Philosophy 591. But my concern, to repeat, is simply with incapacity-based exculpation.

The moral transformation response can make precisely this claim. It can claim that incapacities are morally exculpatory if and only if they derive from morally transformative aetiologies.[80] This claim amounts to a partial answer to the question we'll only tackle in the next chapter, viz what makes an incapacity morally exculpatory. But we can still evaluate this partial answer before tackling the general question. The claim would be false if there are cases in which incapacities are morally exculpatory even in the absence of any morally transformative aetiology.[81]

One dialectical difficulty here is that many exculpatory incapacities are not considered incapacities *at all* absent some obviously salient aetiology. Take an incapacity to distinguish right from wrong. Defendants born of so-called 'rotten social backgrounds' are rarely considered to be genuinely incapable of such moral discernment, while there are fewer doubts about those suffering mental disorders. Regardless of whether those doubts are correct, this highlights that judgements of incapacity are often intertwined with aetiological judgements.[82] That is, questions about incapacities and aetiologies are often bundled together. As a result, it's difficult to disentangle which is doing the work in any intuitive judgment. Perhaps the aetiology is constitutive of an incapacity being exculpatory. But, equally, perhaps the aetiology is merely strong *evidence* that there exists any incapacity at all, an incapacity which would be ipso facto morally exculpatory. Because of this difficulty we will have to assess the moral transformation thesis somewhat indirectly.

Imagine that I fail to perceive any moral reason to account for the interests of my co-worker. I simply cannot understand why harming her would be wrong. As a result, I steal her lunch. Now imagine the cause of my failure of moral perception/understanding was an intervention by a trickster god or a mind-altering machine. Few lists of qualifying aetiologies would include either of these. Intuitively, my incapacity is nonetheless exculpatory. Thus, if the list of morally transformative aetiologies were restricted to anything like those admitted by the current law, it

[80] More precisely: the incapacities are ipso facto morally exculpatory, or the *ground* of that exculpation, not merely that they cause one to be exculpated for other reasons. Given that all incapacities have *some* aetiology, the moral transformation view is meaningful only if the morally transformative aetiologies constitute a reasonably well-defined subset of all possible aetiologies. To specify that definition now would simply be to offer the moral transformation response's complete candidate answer to the general question of what makes an incapacity exculpatory. Given that we're delaying that enquiry, my evaluation is instead more limited.

[81] By 'any' morally transformative aetiology we mean not simply mental disorders, but the entire kaleidoscope of qualifying aetiologies across all of the incapacity rules.

[82] Partly our response will depend on how we understand capacities in general, which, to repeat, we're only tackling in the next chapter.

would fail to account for these exculpatory incapacities which lack a qualifying aetiology. Of course, there's a perfectly innocent explanation for why these aetiologies are omitted: trickster gods and superhuman manipulators don't exist. If they did, they'd get added to the list of qualifying aetiologies.

But we need to be cautious with this response. We could always claim of *any* intuitively morally exculpatory incapacity, post- and ad-hoc, that it thereby must derive from a morally transformative aetiology. But that would make the view unfalsifiable. This doesn't mean that it's wrong. But it does mean that the specification of the qualifying aetiologies must be much broader than anything like the current law. I'll return to a sophisticated version of this claim in the next chapter.[83] But for now, we'll have to consider more indirect ways to evaluate the truth of the moral transformation response. I'll note two such ways.

First, consider other incapacity rules. Relativisations and counterfactual relativisations only rarely and haphazardly impose aetiology requirements.[84] This is evidence that in most cases the incapacities they account for are morally exculpatory regardless of their aetiology. This evidence is not definitive. There may be supervening reasons why relativisations don't or shouldn't track morally transformative aetiologies. Or the law might simply be wrong; perhaps it ought to add aetiology requirements to relativisations. But the simplest explanation for why the law omits aetiology requirements for relativisations is just that the incapacities alone are exculpatory. Their aetiology is not morally transformative.

Second, consider two historical examples in which incapacities and aetiologies come apart. One: ancient and medieval attitudes to 'madness'. They had few true beliefs about the aetiologies of 'madness'. If anything, the dominant explanations referred to *culpable* aetiologies, such as that madness was caused by supernatural punishment for past culpable transgressions. The moral transformation view would therefore predict that madness, in the absence of any (non-culpable) aetiological explanation, would not be deemed exculpatory. But that was not the case. Madness was consistently deemed to be exculpatory.[85] This implies that the aetiology

---

[83] Ch 5 §2.

[84] Two examples: (1) duress (and its relativised standard of reasonable firmness) is negatived by prior association with the duressors (*Z*/*Hasan* (fn 39) [39]); (2) the fear of violence in loss of control (a counterfactual relativisation) is negatived if one incited the cause of that fear to provide an excuse (Coroners and Justice Act 2009 s55(6)(a)). Both aetiology restrictions are explicable in terms of prior fault, discussed below. But most other relativisations are not so limited, except by additional prior fault doctrines like intoxication, also considered below.

[85] See ch 3.

of madness is not driving those judgments. Two: cases where aberrant behaviour was once considered culpable, before later being considered the product of incapacity, and as a result being considered morally exculpated. (As no incapacity was originally recognised, a fortiori no aetiology of that incapacity was recognised.) Examples include the behaviour of dyslexics, Tourette's sufferers, schizophrenics, and others. Did the mere discovery of an incapacity shift attitudes, or was further aetiological information also necessary? My impression is that the mere fact of incapacity did the work. The best evidence for this is simply that the aetiologies of these and many other mental disorders *remain* poorly understood. The dominant paradigm for identifying and taxonomising mental disorders is symptom-based, not aetiology-based.[86] If we don't understand the aetiologies of such conditions, their precise aetiology can't be doing the moral legwork. The mere fact of incapacity is the better candidate.[87]

An objection: both 'madness' and 'mental disorder' are themselves aetiologies for incapacities. That we don't understand the *additional* aetiologies of these conditions is beside the point. After all, the law (of insanity, at least) doesn't add additional aetiology requirements. It only adds first-order aetiology requirements, such as showing that incapacities arise from mental disorders. My examples do nothing to dissuade us that first-order aetiology requirements are unjustifiable.

But this objection misses the mark. My examples are not intended to demonstrate that mental disorders are unjustifiable qua qualifying aetiologies. Rather, they're intended to demonstrate that mental disorders are unjustifiable qua *morally transformative* aetiologies. To be morally transformative, an aetiology must pick out some underlying moral difference-maker. But this would require that the aetiology selected picks out some stable and coherent referent. My point is that 'madness', 'mental disorder', or even 'recognised psychiatric condition', rarely correspond to any such stable referent. This is a consequence of using a symptom-based taxonomy of mental disorder.[88] The aetiology 'mental disorder' might matter morally. It must,

---

[86] The paradigm taxonomy is that of the APA's *Diagnostic and Statistical Manual* (DSM) and the WHO's *International Classification of Diseases* (ICD).

[87] Does this imply that *all* otherwise inexplicable wrongdoing, or even all human conduct, is in principle amenable to causal or mechanistic explanations? If so, does it follow that all wrongdoing, once mechanistically explained, will be deemed non-culpable? I'll have more to say about this common form of free-will scepticism in the next chapter, but it suffices to say here that what I say does not commit me to the additional premises and inferences required to reach those sceptical conclusions.

[88] So much the worse for that taxonomy? Perhaps. The symptom-based paradigm's critics include the former director of the (US) National Institute of Mental Health: Thomas Insel and Philip Wang, 'Rethinking Mental Illness' (2010) 303 JAMA 1970. But critics of this paradigm rarely go so far as to claim that those currently diagnosed as disordered are not thereby truly incapable, for want of any

if the law is to be justified. But the lack of a coherent referent, based on a robust aetiological grounding, implies that it does not matter in virtue of constituting a morally transformative aetiology.

One last retort. The lack of a stable referent in many mental disorders is just a consequence of aetiological psychology being an infant discipline. Perhaps our current labels are placeholder theoretical terms. The aim is still for the law to pick out disorders with robust referents, disorders which constitute morally transformative aetiologies.

One last reply. Let's assume that it's true that some aetiologies morally transform incapacities, making them exculpatory. The law must still formulate rules to identify, specify, and calcify these morally transformative aetiologies. The law's formulations of the qualifying aetiologies will be necessarily over- and under-inclusive relative to their target. This is simply a consequence of the nature of rules.[89] And the consequence of this is that the *law's* aetiology requirements, quite aside from the 'true' aetiology requirements, whatever they might be, will be heuristic devices. That those tests are fulfilled would constitute mere evidence that an incapacity is (morally transformative and thereby) exculpatory. My reply to the moral transformation response can, therefore, run via two routes. The first route is the one given at length above. Aetiologies are not morally transformative. It is the underlying incapacities which are the ground of exculpation, and thus the justificatory puzzle raised by aetiology requirements remains outstanding. The second, alternative, route is to claim that the aetiologies picked out by the law's aetiology requirements are not ipso facto morally transformative. Rather, they only *evidence* the underlying genuinely morally transformative aetiologies. But whichever route my reply takes, we still need to provide some further justification of the law's aetiology requirements. According to the second route we need to justify the law's aetiology requirements qua *proxies for the true, morally transformative aetiologies*. According to the first route, we need some other justification. But the most plausible justification is precisely that the law's aetiology requirements *improve the law's proxies for selecting exculpatory incapacities*. On either route, then, our task is to justify proxies for the true basis of exculpation. The routes are different,

aetiological grounding of that label. Rather, their criticism is mostly scientific: that the current classificatory system fails to cut nature at its joints, slowing our understanding and potential remediation. There are more radical critics of the reality of mental disorders in the absence of any specific aetiological mechanism, such as Thomas Szasz, but that is a minority view. At any rate, when it comes to insanity, the law's concept of a 'disease of the mind' is deliberately and explicitly broader (and thus even less-aetiologically grounded) than the current paradigm of 'mental disorder': *Sullivan* [1984] AC 156, 172.

[89] At least, social rules governing human affairs. See the discussion of Schauer, below.

but our task is largely the same. The suggestion that aetiologies morally transform incapacities doesn't in itself justify the aetiology requirements.

3.2 Prior fault

How can adding aetiology requirements improve the law's proxies for exculpatory incapacities if the law already asks for direct proof of those incapacities in step (1)? The answer follows from the point just made: the law's formulations only indirectly track their target. Imagine if the criminal law contained just a single prohibition against 'harmful public wrongs'. Many criminal law theorists would endorse the sentiment.[90] But none would endorse the implementation. The reason is obvious: such an abstract formulation would leave far too much discretion to officials to determine what counts as harmful public wrongs.[91] Perhaps our current taxonomy of homicide, property offences, sexual offences, etc is only justified insofar as these taxa pick out harmful public wrongs. And perhaps this taxonomy leaves regrettable gaps between the taxa. But it will still perform better than the less underinclusive, but far more difficult to apply, one-category taxonomy. Similarly, perhaps formulating the incapacity rules directly and baldly in terms of the relevant underlying morally exculpatory incapacities would leave those rules far too abstract. We might need to add aetiology requirements to make them workable in application, even if this comes at the cost of some under- or over-inclusivity. That's the general strategy. In this subsection, I'll focus on one version of this approach, before widening the lens in the next subsection. The version I'm interested in concerns *prior fault*.

When laying out the puzzle of the aetiology requirements, I quoted the Law Commission:

> the reason for a defence based on…non-responsibility would be the lack of capacity. The background causes of that lack of capacity are not legally relevant...

But the Law Commission added an important caveat:

> (save where… [the background causes] amount to prior fault of the accused)[92]

[90] This is not to endorse the sentiment. More would support something more nuanced, like 'conduct which if criminalised would reduce the incidence of harmful public wrongs'.

[91] Schauer, *Playing by the Rules* (n 48) ch 7; James Edwards, 'Harm Principles' (2014) 20 Legal Theory 253.

[92] Law Commission, *Insanity and Automatism* (n 4) [2.18].

The classic legal example of prior fault is intoxication: I cannot avoid liability, even when I lack the requisite mens rea for certain crimes, if I would have had the requisite mens rea but for getting drunk.[93] That's because the law considers getting drunk (to that degree) a form of prior fault which can substitute for (proof of) the usual (contemporaneous) fault element, the mens rea.[94]

Similarly, in many circumstances, a defendant may be at fault for lacking a relevant capacity at the relevant time. That is, the aetiology of their incapacity might itself be faulty. The law might want to deny reliance on these faulty aetiologies. It could do so with prior fault rules: by barring the use of incapacity rules if the incapacity arose due to a specified faulty aetiology, like voluntary intoxication. But perhaps the range of faulty aetiologies would be very wide or identifying them would be very difficult. The law might then change tack. Rather than create a *blacklist* of faulty aetiologies, it could instead create a *whitelist* of *non*-faulty aetiologies. That is, the law might permit the use of incapacity rules only if a defendant could prove that their incapacity derived from a faultless whitelisted aetiology. This is exactly how the incapacity rules operate: defendants must prove that their incapacity arose from specific non-faulty aetiologies like young age, a disease of the mind, recognised medical disorders, and so on. Call this is the whitelisting prior fault (*WPF*) approach.

One problem with the WPF approach is that the current aetiology requirements don't capture all or even the bulk of possible non-faulty aetiologies. Admittedly, it's hard to vindicate this claim without a clearer idea about what might constitute an exculpatory incapacity in general, and that discussion has to wait until the next chapter. But it seems plausible that precisely the same underlying relevant incapacities—to understand right from wrong, say—could arise by a wide variety of aetiologies. The classic go-to example is a difficult or traumatic upbringing, the so-called 'rotten social background' claim. Equally, the exact aetiology of a defendant's

[93] Crimes of 'basic intent': *Majewski* [1977] AC 443, *Richardson* [1999] 1 Cr App R 392. (Though is it probably more accurate to talk not of *crimes*, but of *mens rea elements* of basic intent, per *Heard* [2007] 3 WLR 475).

[94] The case law is ambiguous as to whether intoxication just *is* the fault element (albeit varied from crime's standard element) or an alternative to a fault element (warranted by policy concerns). On the one hand, early cases (*Majewski* (ibid), *Kingston* (n 8)) brusquely implied the latter view. On the other hand, some later cases seemingly required that the actual fault element would have been fulfilled counterfactually (so that D would not be reckless, even if they would have foreseen a risk of the proscribed outcome, if that risk was reasonable to run: *Richardson* (ibid)), while some commentators favour the view that *actual* foresight that taking the intoxicant might lead to the type of offending in question: Findlay Stark, 'Prior Fault' (2014) 73 Cam LJ 8; Quratulain Jahangir, JJ Child, and Hans Crombag, 'Prior fault and contrived criminal defences: coming to the law with clean hands' (2017) Institute of Law Review 1.

incapacity might be entirely unknown or even unknowable. Insofar as the aetiology can be disentangled from the incapacity, I see no reason to believe that the law must or does track all possible faultless aetiologies. As such, whitelisting specific non-faulty aetiologies entails denying legal exculpation to some defendants whose non-faulty aetiology for whatever reason didn't make it on to the whitelist. This predictably denies legal exculpation to defendants who are morally exculpated on the basis of incapacity, defendants who are owed some form of legal exculpation in virtue of a culpability principle. That is a problem for the WPF approach.

Another problem with the WPF approach is implicit in something I noted above: that the insanity rules do not add any *additional* aetiology requirements. The insanity defence is available regardless of the aetiology of the disease of the mind. Including, that is, if the disease of the mind was itself culpably induced, eg by voluntary intoxication.[95] In these cases, the disease of the mind is *not* a faultless aetiology. Thus: the WPF approach cannot explain why a disease of the mind qualifies as a whitelisted aetiology given that it can, in fact, be faulty.

An advocate of the WPF approach could simply recommend reform. Like the Law Commission, she could deny that insanity exculpates if culpably induced. Indeed, that answer has the support not only of the Law Commission but also has a clear harmonising rationale. The insanity defence is somewhat anomalous in not restricting its application to non-faulty aetiologies. All of diminished responsibility,[96] loss of control,[97] duress,[98] and self-defence[99] contain prior fault provisions. While the law could go further in harmonising and rationalising these doctrines, this nearly consistent denial of exculpation in cases of prior fault seems like a good place to start to justify and explain the aetiology requirements.[100] For all these doctrines

[95] *Coley v The Queen* [2013] EWCA Crim 223. Contrast automatism: *Bailey*, *Hardie* (n 36). I discussed the rarity of the latter cases above under the label of an *anti-imputation presumption*.

[96] Diminished responsibility is denied if caused by voluntary intoxication*: Dowds* [2012] EWCA Crim 281.

[97] Loss of control is denied for (faultily) inciting the provocative trigger in order to provide an excuse. Criminal Justice Act 2009 ss54(4), 55(6).

[98] Duress is denied for (faulty) prior association with one's duressors. See *Z*/*Hasan* (fn 39).

[99] Self-defence is denied when one relied on facts as one mistakenly believed them to be if that mistake was attributable to voluntary intoxication. Criminal Justice and Immigration Act 2008 s76(5). That is, this is an exception to self-defence's absence of any standard of reasonable belief, which constitutes a counterfactual relativisation: ch 3 §4.2.

[100] For some reform suggestions, see Jahangir, Child, and Crombag, 'Prior fault' (n 94); Stark, 'Prior Fault' (n 94).

blacklist certain faulty aetiologies. The newly reformed and harmonised incapacity rules would ask:

1) Is there a relevant incapacity?
2) Does it arise from a qualifying aetiology?
3) Is that aetiology disqualified in virtue of prior fault?

(Eg, for insanity, step (2) would ask about a disease of the mind.) But think about what this entails. The entire justification of aetiology requirements, step (2), was supposed to be that some aetiologies did (not) disclose prior fault. But, if harmonised, the law would explicitly and separately ask about prior fault. Given that, it may as well cut to the chase and ask

1) Is there a relevant incapacity?
3*) Is that incapacity disqualified in virtue of prior fault?

The aetiology requirement becomes redundant. By harmonising with the current law, we might justify *blacklisting* certain faulty aetiologies, but we would struggle to justify the proposed *whitelisting* prior fault approach.

It's true that under some conditions the blacklist and whitelist would be extensionally identical. (Meaning if one is justified the other would be too). But those conditions include perfect information as to what counts as an exculpatory incapacity (conditional on aetiology), perfect information as to who fulfilled those criteria, and perfect application of that perfect information to individual cases. Under those conditions, the blacklist would include all and only faulty aetiologies, and the whitelist would include all and only non-faulty aetiologies. But those conditions don't obtain. The choice of approach is, therefore, in part, a choice between varieties of erroneous application.[101] Using a blacklist favours false exculpation, presuming by default that incapacities have non-faulty aetiologies. Using a whitelist favours false inculpation, presuming by default that incapacities have faulty aetiologies. Seen this way, it's clear why the criminal law typically accounts for prior fault by blacklisting those faulty aetiologies. This fits naturally with the presumption of innocence and the mens rea principle, the principle that the prosecution must prove mens rea in core criminal offences.[102] The intoxication rules, for example, blacklist one specific aetiology for criminal wrongdoing: intoxication. To justify the

[101] Aka type I and type II errors.

[102] *Woolmington v DPP* [1935] AC 462.

aetiology requirements as a means of accounting for prior fault on the basis of doctrinal harmonisation would require that those other rules relating to prior fault switch from a blacklist to a whitelist approach. No-one advocates that regarding the general intoxication rules: most commentators recommend a *less* strict approach. A common suggestion is to *add* a requirement that defendants were reckless as to the risk of future offending when getting drunk.[103]

But some commentators do recommend the kernel of a view about defences which might help to justify the WPF approach.[104] The core idea is that defences are morally discretionary and may justifiably be restricted to account for various policy concerns. Insofar as incapacity rules are defences, then, this explains why we should start, by default, with a presumption against legal exculpation on the basis of incapacity. From that starting point, we can then add a whitelist of permissible non-faulty aetiologies. But we are not obligated to exculpate all non-faulty aetiologies. We therefore face no justificatory burden in preferring a whitelist approach which favours false inculpation to a blacklist approach which favours false exculpation.

This strict approach is not a thesis about prior fault, incapacities, or aetiologies. It's a thesis about the permissible denial of (excusatory) defences. It is the thesis that defendants rarely have a moral right to an excusatory defence,[105] and that instead the law has a broad moral permission to not grant them.[106] It leaves open what might justify granting excuses: being morally exculpated on the basis of incapacity derived from a faultless aetiology is but one candidate. But it does operate to block objections that the law's excuses are under-inclusive, ie

[103] Jahangir, Child, and Crombag, 'Prior fault' (n 94); Stark, 'Prior Fault' (n 94).

[104] Jahangir, Child, and Crombag claim that different rules ought to apply for prior fault as an inculpating factor as opposed to prior fault used to deny an otherwise exculpating factor. The argument for this view is unclear. Mark Dsouza offers an argument: intoxication qua inculpating factor caps liability at basic intent crimes, whereas intoxication qua denial-of-exculpating-factor (eg denying self-defence) leaves open the full degree of criminal liability. From this we might infer that the latter are too harsh, per those in the footnote above. This makes it all the more curious that Dsouza favours a stricter approach for *non*-faulty unreasonable beliefs in self-defence, as discussed in the text below. Mark Dsouza, 'Intoxication, psychoses, and self-defence: Evaluating *Taj* [2018] EWCA Crim 1743' (2018) Arch Rev.

[105] That is, to the *creation or existence* of a defence covering one's circumstances; a moral right may arise to the *use* of a pre-existing defence if one's circumstances meet the qualifying criteria.

[106] This permission is asymmetrical. The law can't have an unfettered moral permission to grant excuses; that would permit unconscionable excuses on the grounds of political connections, religion, ethnicity, etc.

that they permit excessive false inculpation. By asymmetrically favouring under-inclusivity over over-inclusivity, it thereby favours the whitelist over the backlist approach.

This strict approach is a thesis about excuses, and not all incapacity rules operate by way of excuses.[107] But I want to leave open whether that ought to be the case. So it's worth tackling the strict approach head-on. It's not usually spelled out like this. But it underlies various approaches to excuses, including to those incapacity rules that (part-) constitute excuses. Mark Dsouza, for example, would deny self-defence not only to those who unreasonably believe in a threat due to intoxication but, in the name of public protection, also if due to (non-culpably induced) mental disorder.[108] William Wilson would require that all excuses derive from a reaction to an 'external crisis', and would thus deny exculpation to those whose incapacities are purely internal.[109] And Jeremy Horder would deny any excuse if doing so would conflict with the law's 'strategic aims'.[110] Each of these suggestions implies that there are strong general reasons to deny excusatory defences even for non-culpable defendants. But this approach is hard to justify. Take Dsouza's suggestion to bar reliance on unreasonable beliefs in self-defence if they arise due to mental disorders. Perhaps this would better protect the public from disordered mistaken self-defenders. Sure. But why single out the mentally disordered? Perhaps this is a proxy for those more likely to re-offend.[111] But it's at best a weak proxy and comes with the large downside of picking out one (often faultless) aetiology for special (worse) treatment by the criminal law. Surely it would be more robust and simpler, not to say fairer, simply to impose an objective standard of reasonable belief in a threat. Voila: no more reliance on unreasonable beliefs in self-defence, of whatever origin. But the law has not taken that step, despite ample opportunity. That's because the test of genuine belief is a counterfactual incapacity relativisation. The law is committed to accounting for incapacities, despite the public protection downsides. (Which could be ameliorated through non-criminal means). The law's implicit culpability principle rules out that strict approach to excuses and with it the WPF approach.

---

[107] See ch 3.

[108] Dsouza, 'Evaluating Taj' (n 104) 9

[109] William Wilson, 'The Filtering Role of Crisis in the Constitution of Criminal Excuses' (2004) 17 Can JL Juris 387.

[110] Jeremy Horder, *Excusing Crime* (Oxford 2004) ch 1.

[111] Ie, commit the offence—regardless of whether self-defence is available as a defence.

Much the same can be said of Horder. He claims that the agent-centred requirement for culpability is a necessary but insufficient condition for making excuses available. The sufficient condition is congruence with the law's strategic aims, which include: a match between how prohibitions ought to be and are regarded; a requirement that defences do not undermine law-abidingness; and discouragement of a 'defence industry' for 'unmeritorious claims'.[112] He claims that '[e]ven a mere chance that such an undermining of respect [for law] may be the result…could be enough to justify refusing to [offer a defence].'[113] All of this is entirely at odds with the demands of a culpability principle. It's one thing to convict some non-culpable defendants due to unavoidable evidential standards or to prevent excessive perverse incentives. It is quite another to deny defences to non-culpable defendants for the mere chance that doing so undermines others' respect for the law. That surely isn't a weighty enough consideration to justify criminalising non-culpable conduct types. And if we reject this strict approach to defences in general, we have very little basis to accept the WPF approach, and with it the aetiology requirements.

My argument has not been that prior fault does nothing to justify the aetiology requirements. It is sometimes justified to deny reliance on an incapacity rule if the incapacity was culpably induced. But the most sensible implementation of that idea would be to blacklist those faulty aetiologies. Not to whitelist only some non-faulty aetiologies. At best, prior fault can supply one element within a broader set of considerations that together justify the aetiology requirements as justified proxies. I turn to those other considerations now.

3.3 Error reduction

The broad claim under consideration is that, appearances to the contrary notwithstanding, the aetiology requirements really do improve the law's proxy for culpability. As we saw in the previous section, to do this it must achieve a more valuable mix of under- and overinclusivity. The main way of achieving that value is to increase accuracy between the law's *rationale* and its *application* to a set of cases in practice. The ultimate rationale for legal rules, I'll assume, is achieving value, however specified. The narrower rationale we're interested in is avoiding convicting disproportionately to culpability, our culpability principle. And the specific rationale for having the incapacity rules is avoiding convicting disproportionately to lowered culpability on the basis of incapacity. Given that specific rationale, the incapacity rules would be under-

[112] Horder, *Excusing Crime* (n 110) ch 1.1-1.2.

[113] ibid 18.

inclusive if they failed to legally exculpate those who are relevantly incapable, and overinclusive if they exculpated those who are not relevantly incapable. The aetiology requirements would improve the law's accuracy, its proxy for culpability, if they reduced these two types of error. Do they?

If the aetiology requirements were removed, then courts would have to ask directly and explicitly whether a defendant was less culpable in virtue of a relevant incapacity. Prima facie, this would mean the law would account for more incapacities than the currently aetiology-restricted rules. Prima facie, then, removing the aetiology requirements would reduce the law's underinclusivity. But, again prima facie, this would come at the cost of overinclusivity. Liberalising the incapacity rules by removing the aetiology requirements would open up a wide range of new defence strategies based on novel incapacity claims. Faced with such claims, law-appliers might fail properly to filter out unmeritorious claims, resulting in overinclusivity in application. If this occurred, and its disvalue was worse than the disvalue of underinclusivity, then the error-reduction answer might justify the aetiology requirements.

I mentioned a lot of 'prima facies'. That's because what would actually happen if the aetiology requirements were removed is necessarily speculative. But one plausible scenario if the aetiology requirements were removed is backlash. Lawmakers would baulk at the newly overinclusive incapacity rules or their application to particular cases. In response, they might switch to stricter rules across the board. They might eliminate (counterfactual) relativisations in favour of non-relativised standards or eliminate certain incapacity doctrines altogether.[114] Relative to that counterfactual, the law's aetiology requirements might avoid not just overinclusivity, but also underinclusivity. Under that counterfactual they seem clearly justified, improving both sides of the equation. But selecting the appropriate counterfactual is a tricky business. Perhaps after implementing those stricter rules, there would be a counter-backlash regarding cases where clearly non-culpable conduct was not legally exculpated. The law might be reformed again to account for more incapacities, even compared with the current law.[115] Relative to this new counterfactual, the law's aetiology requirements might, once again, seem unduly underinclusive. We can go back and forth. Ultimately, I think all counterfactuals are relevant. But trying to account for all of them is methodologically intractable. On the other hand, we can't just use an arbitrary cut-off in the space of possible counterfactuals (temporal

[114] This plausibly explains the abolition of the doli incapax doctrine, and the reintroduction of a standard of reasonable belief in sexual offences in the Sexual Offences Act 2003 following *Morgan* (n 26).

[115] The process might be something like my lightly fictionalised legal fable when introducing counterfactual relativisations in ch 3 §1.3.

or otherwise). I'll therefore limit the analysis to the counterfactual where the rules under consideration are implemented, rather than considering how those very rules might change as a result.[116] So let's start over. Removing the aetiology requirements would (in expectation) increase overinclusivity but decrease underinclusivity. Is the trade worth it?

I've already rehearsed why underinclusivity is disvaluable. It entails failing to give a break to those who are morally exculpated, violating a culpability principle. The case of infancy is instructive. Infancy uses young age *only* as a proxy, not as an aetiology requirement on top of demanding proof of the relevant incapacities. That makes it especially prone to overinclusivity: there's no way to tell if any particular child falls within the rule's rationale; that is, whether they are relevantly incapable. But, assuming my argument against Yaffe is right, the law simply accepts that overinclusivity. Why? Presumably because that overinclusivity is a fair price to pay to avoid underinclusivity. Better to let off the occasional capable child than to haul hundreds of incapable children before the courts.[117] Indeed: better to let off one capable child than to criminalise one incapable child. The criminal law has an inbuilt preference for underinclusivity over overinclusivity, as immortalised by Blackstone's ratio. The law's approach to infancy reveals that preference. In doing so it heightens the justificatory puzzle presented by the aetiology requirements. Incapacity rules with these requirements already avoid overinclusivity by asking for direct proof of the relevant incapacity. The only source of overinclusivity that remains is when the law's formulations of the relevant incapacities are overinclusive or else those formulations are applied overinclusively. Can avoiding that really be so valuable as to outweigh the obvious disvalue of underinclusivity?

That depends. Overinclusive formulations in attempting to account for the relevant incapacities are not just the result of obvious errors or even (as the next chapter explores) the genuine difficulty of specifying the nature of the relevant incapacities.[118] Overinclusivity is also

---

[116] This is I think the default methodological position when considering rules changes. The difficulty is analogous to the epistemic problem for consequentialism (eg Tyler Cowen, 'The Epistemic Problem Does Not Refute Consequentialism' (2006) 18 Utilitas 383) and considerations of ideal/non-ideal theory (eg Laura Valentini, 'Ideal vs. Non-ideal Theory: A Conceptual Map' (2012) 7 Phil Com 654).

[117] Some say that the variable pace of child development makes bright lines inapt. Heather Keating, 'The "responsibility" of children in the criminal law' (2007) 19 Child Fam LQ 183. In English medical law the test of a child's competence is based (in theory) on the individual child's actual capacities of understanding rather than presuming based on age. *Gillick v West Norfolk and Wisbech AHA* (1985) 3 All ER 402 (HL).

[118] One obviously over-inclusive formula is insanity's wrongness limb, which according to Windle requires ignorance of legal wrong. This is obviously over-inclusive as a defendant who falsely believes that honour killings are legally permissible falls within that formula even if the defendant correctly

simply in the nature of rules and of language. As Frederick Schauer has pointed out, if a decision-maker deviated from a rule whenever it turned out to diverge from the background rationale, then that decision-maker is *not* making rule-based decisions at all. They are only making rule-based decisions if they would follow the rule *despite* it diverging from the background rationale (in their estimation).[119] Why bother with rules then? The answer is necessarily context dependent. Sometimes rule-based decision-making *won't* have anything going for it.[120] But the most common and most powerful justification for rule-based decision-making is that it provides a better practical decision procedure than if decision-makers relied on the background rationale alone.[121] Judges might often get it wrong as to who counted as relevantly incapable.[122] The subjects of the law might accordingly find it hard to rely on those judgments.[123] Rule-based decision-making can avoid these problems by solving matters of coordination and more appropriately allocating decisional jurisdiction.[124] By generalising across cases, rule-based decision-making frees judges from considering every nuance of every case.[125] Well-formulated rules improve overall decision-making. All these generic benefits of rule-based decision-making help to justify the use of aetiology requirements rather than simply leaving the specification of the relevant incapacities in the hands of judges.

This is the most common justification offered for the aetiology requirements. Richard Posner claims that the insanity defence is economically justifiable as the insane are not deterrable. But, he continues, the 'reason for requiring proof of mental disease, and not just proof of undeterrability, is to focus inquiry and to reduce the risk of a legal error in the defendant's favor.'[126] Paul Robinson claims that the law is concerned with the '*cause* of the excusing

---

believes it to be morally impermissible, and yet does it anyway. That defendant seems clearly culpable. James Manwaring, 'Windle Revisited' [2018] Crim LR 987, and 'The Wrongness Limb' (draft).

[119] Schauer, *Playing by the Rules* (n 48) 100-102.

[120] ibid ch 7.

[121] ibid 135.

[122] ibid 149ff.

[123] ibid 137-145. John Gardner claims that excusatory defences ought not to be considered when deciding how to act: to calculate that one might benefit from a provocation defence, for example, would be to undermine the very basis of that defence. John Gardner, 'The Gist of Excuses' in *Offences and Defences* (Oxford 2007) 138. Granting this is true (arguendo), it might remain problematic that third parties cannot work out how the rules apply to a particular case (eg in working out whether to intervene, or in more diffuse contexts like insurance claims).

[124] Schauer, *Playing by the Rules* (n 48) 143-145; 158-166.

[125] ibid 31-35; ch 2-5.

[126] Richard Posner, 'An Economic Theory of the Criminal Law' (1985) Columbia Law Review 1193, 1223-1225.

conditions' as that amounts to an 'independently observable phenomenon'.[127] Presumably, that is, in a way that the underlying incapacities are not—all the better, then, for rule-based decision-making.[128] HLA Hart notes that asking (as the German system does)

> whether the accused "lacked the ability to recognise the wrongness of his conduct and to act in accordance with that recognition" […] raises formidable difficulties of proof, especially before juries. For this reason I think that, instead of a close determination of such questions of capacity, the apparently coarser-grained technique of exempting persons from liability to punishment if they fall into certain recognized categories of mental disorder is likely to be increasingly used. Such exemption by general category is a technique long known to English law; for in the case of very young children it has made no attempt to determine, as a condition of liability, the question whether on account of their immaturity they could have understood what the law required and could have conformed to its requirements, or whether their responsibility on account of their immaturity was 'substantially impaired', but exempts them from liability for punishment if under a specified age. It seems likely that exemption by medical category rather than by individualized findings of absent or diminished capacity will be found more likely to lead in practice to satisfactory results...[129]

These points are well taken. But they don't say enough to justify the current aetiology requirements. They suggest one reason in favour of clear rules picking out aetiologies. But they don't attempt to balance that with the reasons against doing so.[130] The contrary argument is that there are a variety of more sensitive means to avoid overinclusivity without sacrificing so much underinclusivity.

Evidential presumptions are one means of avoiding overinclusivity. It would be difficult for the prosecution accurately to prove a defendants' lack of capacity. But the law has a solution:

---

[127] Paul Robinson, *The Structure and Function of Criminal Law* (Oxford 1997) 92. A similar point is emphasised by Loughnan, *Manifest Madness* 24-25, who claims that a focus on disabilities, and not their results emphasises the objective aspects of incapacity rules and the role of expert medical knowledge. [But is that disabilities or aetiologies?]

[128] In fairness to Robinson, he is concerned that the current law fails to excuse where multiple aetiologies cumulatively cause a relevant incapacity (though he thinks that relativisations go some way to rectifying the situation): Robinson, *Structure and Function* (n 127) 93-94.

[129] HLA Hart, *Punishment and Responsibility* (John Gardner ed, 2nd edn, Oxford 2008) 228-229.

[130] Category-based decision-making is controversial in other areas too. In the context of consent to offences against the person, for example, some argue that the current category-based approach fails to do justice to autonomy (the underlying background rationale): Julia Tolmie, 'Consent to Harmful Assaults the Case for Moving Away from Category Based Decision Making' [2012] Crim LR 656.

it imposes a *presumption of capacity*. In the *M'Naghten* formulation, 'every man is to be presumed to be sane, and to possess a sufficient degree of reason to be responsible for his crimes, until the contrary be proved…'. This means the law errs towards underinclusivity. Some find that problematic. The Supreme Court of Canada, for example, accepted that the presumption of capacity conflicts with the (constitutionally mandated) presumption of innocence. Their reasoning, in short, was that some defendants who would be found not guilty as a result of their incapacity might yet fail to rebut the presumption, and thus be found guilty without the prosecution proving as much. Still, the court accepted that the presumption of capacity was nonetheless justified on the basis that it would be an 'impossibly onerous burden' for the crown to disprove incapacity.[131] In other words, the presumption of capacity avoids overinclusive incapacity rules at the cost of some underinclusivity. That cost was held to be justified as a proportional limitation on defendants' rights. It is proportionate, in part, as defendants can nonetheless rebut that presumption on the balance of probabilities. But it would not be justified if defendants had to rebut the presumption of capacity beyond a reasonable doubt, still less if the presumption was irrebuttable.[132] If the Canadian Supreme Court is right, the current presumption of capacity might be required to avoid overinclusivity. But it takes the law to the borderline of acceptable limitations on defendants' rights. The aetiology requirements impose an *additional* hurdle for defendants. This risks taking the law over that line, making its exculpation unjustifiably underinclusive.

Abolishing the aetiology requirements would avoid excessive underinclusivity. If that proved too overinclusive, the law could leave it to the *prosecution* to prove that the defendants' incapacity ought to be disqualified from consideration on the basis of its aetiology. Or it could impose a merely evidential burden of proof on the defendant to rebut a presumption of capacity.[133] The current aetiology requirements are a strangely blunt tool given the care that the law takes in qualifying the presumption of capacity.

Still: maybe the incapacity rules would remain overinclusive under those regimes. Maybe further limitations are reasonable. But the law still needn't jump straight to its current aetiology requirements. As we saw above, the current aetiology requirements amount to a presumption

[131] *Chaulk* [1990] 3 SCR 303. (This is not to agree with the courts' interpretation of either the presumption of innocence or capacity).

[132] They noted that the legislature was not required to pick the *best* possible derogation from the presumption of innocence, and consider some more complex evidential arrangements. But I see no evidence that they would accept such a high evidential burden on the defendant.

[133] The defendants in *Chaulk* (n 131) favoured this approach, as did Wilson J's dissent.

that defendants' incapacities are non-exculpatory in virtue of their aetiology. That 'presumption' can only be rebutted by proof that the aetiology fell within a qualifying category (mental illness, young age, etc). Aetiologies that fall outside of the qualifying categories, by contrast, are simply irrelevant. That sharp cut-off leaves the aetiology requirements underinclusive. Instead of this sharp cut off, the law could instead simply raise the standard of proof for defendants who would wish to rely on morally exculpatory aetiologies outside of the law's standard categories.[134] That would deal with concerns of overinclusivity without simply barring defendants from reliance on incapacity rules if their incapacity happened not (provably) to derive from the law's pre-selected qualifying aetiologies. It seems unlikely that the law's current approach to the aetiology requirements is an acceptable solution to the problem of error.

To recap, the law's options include:

1) No aetiology requirements
2) Burden on the prosecution to disqualify an aetiology (to various proof standards)
3) Burden on the defendant to qualify an aetiology (to various proof standards)
4) Aetiology requirements

Given that the presumption of capacity already goes a long way to reducing the risk of overinclusive incapacity rules, using the current aetiology requirements, option (4), seems like the approach least likely to be justified.

Alongside evidential presumptions, the law could achieve a better mix of under- and overinclusivity by modifying its substantive rules and exceptions. Rather than containing a fixed and exhaustive list of qualifying aetiologies, the law could add an additional generic category to catch nonstandard aetiologies; the familiar technique of the ejusdem generis clause. This clause could itself have exclusions, eg to account for prior fault. And it could be subject to prosecution-friendly evidential standards of the sort just discussed. Such a clause would avoid the stark underinclusivity of the current aetiology requirements without risking the overinclusivity that may result from outright abolition.

These suggestions—altering evidential standards and adding substantively to the list of permissible aetiologies—do come at the cost of added complexity. It's possible that this

[134] I would prefer a tailored standard of proof, eg as suggested by Gustavo Ribeiro, 'The Case for Varying Standards of Proof' (manuscript 2016), but if current practice is followed this would amount to requiring the defendant to prove that her aetiology qualified as exculpatory beyond a reasonable doubt.

complexity might mean the incapacity rules (and procedures) are mistakenly applied, resulting in a worse fit with their rationale. Or, more likely, such complexity might come at the cost of other values in the criminal justice system, like affordability, simplicity per se, and so on. But, once again, these costs of complexity must be weighed against the benefits. The primary benefit is not convicting non-culpable defendants, a benefit to which the criminal law attaches a high opportunity cost.

But let's say the cost of complexity is too high. What follows? Well, the simplest solution would be to abolish the aetiology requirements altogether. That is both simpler than requiring proof of (often complex) aetiologies and, given the arguments above, arguably better conforms to the rules' rationale. Of course, abolition too has its potential downsides, including a failure to label the source of incapacity. My point is not that abolishing the aetiology requirements is obviously the correct move, nor even that the aetiology requirements could not be justified. My point is only that, given the range of reasons in play, and the range of possible legal responses, the aetiology requirements seem like one of the worse solutions to the problem of error.

Conclusion

This chapter considered how to bridge the gap between moral exculpation and legal exculpation. There is no way of closing it completely. That is in the nature of rules. I have argued that we are justified in having some incapacity rules rather than none and that they can, at their best, constitute a good proxy for lowered culpability.

But I have doubted whether the aetiology requirements can be justified. They appear to leave the law's exculpatory incapacity rules too underinclusive. The justifications offered in their defence fall short of the mark. This doesn't mean that no such justification is possible. The final justification that I considered—that they achieve a better mix of over- and underinclusivity than rival rules—is necessarily context-dependent. We won't know if it ultimately succeeds or fails (or how well or badly it fares) until we have a sense of the underlying numbers on each side of the equation. We need to know the extent of exculpatory incapacities and the difficulty of proving their (non)existence, to know the extent to which the aetiology requirements render the incapacity rules underinclusive. In short, we need an account of which and why incapacities exculpate and how we can come to know it. Only with those questions answered can we reassess this prima facie case against the aetiology requirements. It is to that task we turn in the next chapter.

5 Why Incapacities Exculpate

I've spent a long time explaining and justifying the incapacity rules, on the assumption that incapacities exculpate. But is that assumption warranted? Why do incapacities exculpate?

I start by asking *which* incapacities exculpate. Three incapacities are usually cited: normative, epistemic, and volitional (§1). Next, I consider what explains why these incapacities are exculpatory. I consider three candidate explanations: that incapacities exculpate if they are *abnormal* (§2); if they entail an incapacity *to respond to reasons* (§3); and if they entail an incapacity *to conform* to the relevant norm (§4). I endorse the last view. Finally, I consider how the conformability explanation interacts with the 'ought implies can' principle (§5) and certain views about free will (§6).

1 Three incapacities

Incapacities exculpate if they negate culpability. Incapacities are not the only culpability-negaters. The list of culpability-negaters also standardly includes the justifications and excuses: self-defence, duress, necessity, and others. Why do any of these factors exculpate? Sometimes, especially in legal argument, their exculpatory nature is taken to be self-evident or sufficiently evidenced by a long history of legal recognition. Philosophers tend to be less accepting of that stance. They typically seek to uncover further factors in virtue of which the items on the list exculpate. Philosophical accounts often prove useful when attempting to resolve controversies about the practical application of these doctrines. For example, we might need to know *why* self-defence exculpates in order to answer whether *mistaken* self-defence exculpates. But there are fewer attempts to explain why incapacities exculpate. As with the lawyers' response to other exculpatory factors, it is often taken to be self-evident that some incapacities exculpate. But, as we'll see, there are controversies about the proper extent of incapacity-based exculpation, just as with other factors. A general account of why incapacities exculpate would help us to resolve these controversies.

Start with *which* incapacities exculpate. HLA Hart's list of relevant incapacities is widely cited and endorsed. He focused on (in)capacities of

> understanding, reasoning, and control of conduct: the ability to understand what conduct legal rules or morality require, to deliberate and reach decisions concerning these requirements, and to conform to decisions when made.[1]

Call these, respectively, *normative*, *epistemic*, and *volitional* incapacities.[2] While formulations differ, the exculpatory relevance of each is widely endorsed.

Normative incapacity lies at the heart of insanity and infancy. It is captured by the M'Naghten rules' famous ability to 'know right from wrong'.[3] Other formulations refer to an ability to *appreciate* wrongfulness[4] or the criminality of one's conduct.[5] These formulations are not extensionally identical. It's controversial whether what matters is bare cognisance of wrongfulness or a more full-blood appreciation.[6] There's also controversy as to which normative domain that wrongfulness is indexed.[7] Is it criminal or legal wrongfulness?[8] Social disapproval?[9] A capacity of critical evaluation?[10] Much recent work links normative capacity to the idea of reasons-responsiveness.[11]

---

[1] HLA Hart, *Punishment and Responsibility* (John Gardner ed, 2dn edn, Oxford 2008) 227.

[2] Victor Tadros similarly refers to 'epistemic, evaluative and volitional' capacities. Victor Tadros, *Criminal Responsibility* (Oxford 2005) 57.

[3] Deriving, recall, from the earlier 'distinguishing good and evil' test, aptly labelled 'moral capacity' by the US Supreme Court in *Clark v Arizona* 548 U.S. 735; 126 S. Ct. 2709 (2006) and a 'capacity for normative understanding' by Eric Colvin, 'Exculpatory Defences in Criminal Law' (1990) 10 OJLS 381, 401. The 'distinguishing right from wrong' formulation was also found in the doli incapax rules.

[4] Law Commission, *Criminal Liability: Insanity and Automatism* (Discussion Paper, 2013) [4.19], and [4.8]ff on the meaning of 'appreciate'.

[5] ALI, *Model Penal Code* (1962) 4.01.

[6] The latter might involve a motivational component. The distinction has parallels with that between reasons externalism and internalism.

[7] A non-controversial case is unfitness to plead, where the relevant capacities refer to the trial process itself, eg 'to comprehend the course of proceedings' *Pritchard* (1836) 7 C&P 303, 304.

[8] For insanity in England and Wales the test is legal wrongfulness, though I have elsewhere criticised the judgment used to reach that conclusion: James Manwaring, 'Windle Revisited' [2018] Crim LR 987.

[9] In Australia the test refers to social morality (the understanding of ordinary reasonable people) (*Stapleton* [1952] HCA 56). In Canada the test is similar (*Chaulk* [1990] 3 SCR 303).

[10] The best answer, I think, is more complicated than simply picking a single domain. I discuss this in the 'The Wrongness Limb' (draft).

[11] Tadros, *Criminal Responsibility* (n 2) 56 refers to the ability to '[respond] to normative reasons'.

Epistemic incapacity features in the M'Naghten rules (knowing the nature of one's action) and in diminished responsibility (understanding the nature of one's conduct).[12] Unlike with normative capacity, most agree that epistemic capacity requires more than bare cognisance of certain facts. Hart referred to deliberation. One textbook describes automatism as the absence of the capacity '*deliberatively* to control one's conduct… [ie being] responsive to a capacity to reason and deliberate about one's conduct'.[13] That is to say, self-control must be exercised via our epistemic capacities of reason and deliberation. Others refer to rationality,[14] reasoning,[15] practical reasoning,[16] and the like.[17] It's controversial whether epistemic (and normative) capacity requires some form of success; that is, whether the agent must form *true* beliefs and evaluations.[18] There is also the question of how normative and epistemic capacity interact. They are often merged in references to a capacity to act intentionally.[19]

Volitional incapacity is absent from the M'Naghten rules but added to the insanity rules in some jurisdictions via 'irresistible impulse' tests.[20] Those tests are highly controversial. There's no controversy that *some* volitional capacities, such as to 'control one's physical actions'[21] or 'exercise self-control'[22], can be relevant to culpability. But they're mainly accounted for via the

---

[12] Insanity per *M'Naghten;* diminished responsibility per Homicide Act 1957 s2(1A)(a). Labelled 'cognitive capacity' by the US Supreme Court in *Clark v Arizona* (n 3).

[13] Simester et al, *Simester and Sullivan* (6th edn, Hart 2016) 112-113 (emphasis in original), endorsed in Law Commission, *Insanity and Automatism* (n 3) [5.2].

[14] Diminished responsibility's capacity 'to form a rational judgement': Homicide Act s2(1A)(b).

[15] Loss of control's 'reasoning faculties': *Oneby* (1727) 92 ER 465, cited by Law Commission, *Insanity and Automatism* (n 3) [3.26].

[16] Law Commission, *Insanity and Automatism* (n 3) refers to 'practical reasoning/rationally forming a judgment'; Anthony Kenny, *Freewill and Responsibility* (Routledge 1978) 79-80.

[17] *Sullivan* [1984] AC 156, 172 describes insanity's 'defect as reason' as referring to capacities of 'reasoning, memory, and understanding', citing Devlin J in *Kemp* [1957] 1 Q.B. 399, 407.

[18] Tadros, *Criminal Responsibility* (n 2) 55-57 says yes, requiring that the agent be capable of 'forming true beliefs and evaluations about the world… [both] in terms of the coherence of his beliefs…[and their] resemblance to the real value of things…'

[19] Justinian's Digest referred to a capacity to form *wrongful* intentions. The Canadian Supreme Court took insanity to involve an incapacity to form a *criminal* intention (*Chaulk* (n 9) 1329). Infancy and doli incapax are sometimes said to invoke an incapacity of intentional action simpliciter (ch 2 §2.2). Elsewhere the Digest refers to a capacity of 'understanding', reflected in Hawkins' reference to an ability to 'understand the law'.

[20] Eg as recommended by the Home Office, *Report of the Committee on Mentally Abnormal Offenders* [Butler Report] (Cmnd 6244, 1975) ch 18.35.

[21] Law Commission, *Insanity and Automatism* (n 3) [4.34]-[4.36].

[22] In diminished responsibility: Homicide Act s2(1A)(c). The capacity to exercise willpower was considered relevant in *Byrne* (1960) 44 Cr App R 246.

requirements for an actus reus, a voluntary act, or a non-automatic act. The controversy refers to the reach of such volitional incapacities and whether they include a motivational component. As with normative and epistemic capacities, volitional capacities have thicker and thinner versions, ranging from mere bodily control, to a capacity to conform one's conduct to the law,[23] to a capacity to form coherent plans of action.[24]

What binds these incapacities together to make them exculpatory? Or is that a leading question? Perhaps the answer is 'nothing'. Perhaps they are but three distinct bases of exculpation. But I don't think that's right. I noted that the lines between these incapacities are often blurred. That would be odd if they picked out three distinct bases of exculpation. Further, these three incapacities are mentioned, in the main, by the incapacity *doctrines*. But what about incapacity relativisations and counterfactual relativisations? (Ie rules where standards are varied or omitted to account for incapacities.) It's rarely specified which incapacities are relevant to (counterfactual) relativisations. That's because there are many plausibly relevant incapacities. Recklessness involves a counterfactual relativisation: it requires inadvertence, not *unreasonable* inadvertence. No standard of belief is imposed in order to account for those who are incapable of adverting to the relevant risk.[25] But that incapacity to advert to a risk might be composed of any number of subsidiary incapacities: visual, auditory, intellectual, rational, even motivational. (D is equally inadvertent of a risk, no matter if her failure was one of hearing, seeing, comprehending, computing, or investigating.) The relevant capacity is simply the capacity *to advert to the risk*. But given the huge variety of potential risks, and potentially inhibitory incapacities, it seems implausible that all relevant incapacities could be slotted comfortably into the 'normative', 'epistemic', or 'volitional' buckets. There must be some unifying explanation for the exculpatory (ir)relevance of any fine-grained incapacity description.

I'll consider three such unifying explanations. First, perhaps incapacities exculpate only if they are in some sense *abnormal*. Call this the *normality thesis*. Second, perhaps incapacities exculpate only insofar as they instantiate a deeper underlying incapacity. Two candidates stand out: an

[23] John Curwood, *Hawkins' A Treatise of the Pleas of the Crown Vol I* (8th edn, Sweet, Maxwell and Stevens 1824) 1-2; ALI, *Model Penal Code* (1962) 4.01; Law Commission, *Insanity and Automatism* (n 3) [3.1]

[24] Tadros, *Criminal Responsibility* (n 2) 56 would require the capacity of 'realising…beliefs and evaluations in action. [Ie]…form[ing] coherent plans of action…'.

[25] Ch 2 §4.1.

incapacity *to respond to reasons* (*Response-ability*), and an incapacity *to conform to the relevant norm* (*Conformability*). I'll defend Conformability against the first two explanations.

2 Normality

Some claim that criminal law speaks specifically to the 'normal individual'.[26] They claim that the abnormal fall outside its prescriptions.[27] Insanity, for instance, is often said to exculpate because insane defendants are abnormal.[28] The old doli incapax doctrine was criticised (and ultimately abolished) as it was said to use a presumption of subnormality.[29] And abnormality has also been the focus of various proposals to reform various incapacity rules.[30] This matches the usage outside of the law. Psychologists talk of 'mental abnormality'. They have a field called

[26] *Kennedy (No 2)* [2007] UKHL 38 [14]: 'The criminal law generally assumes the existence of free will. The law recognises certain exceptions, in the case of the young, those who for any reason are not fully responsible for their actions, and the vulnerable, and it acknowledges situations of duress and necessity, as also of deception and mistake. But, *generally speaking, informed adults of sound mind* are treated as autonomous beings…' The US Supreme Court agrees. In *Morissette v US* 342 US 246 (1952), 250-1 they noted that a 'belief in freedom of the human will and a consequent ability and duty of the *normal individual* to choose between good and evil [is universal and persistent in mature systems of law].' Cited in John Coffee, 'Does "Unlawful" mean "Criminal"?: Reflections on the Disappearing Tort/Crime Distinction in American Law' (1991) 71 Boston University Law Review 193, 210-11. (Emphases added).

[27] This is implicit in Duff's unusually demanding vision of the political community who are the subjects of the criminal law: RA Duff, *Punishment, Communication, and Community* (Oxford 2001) ch 2.

[28] *Coley* [2013] EWCA Crim 223 [16] translates M'Naghten's 'defect of reason' as a 'mental *abnormality*'. A Canadian judgment, *Chaulk* [1990] 3 SCR 303 at 1329 explained that insanity exculpates as the defendant would 'have a frame of reference which is significantly *different* than that which most people share'.

[29] In *C (a Minor) v DPP* [1996] AC 1 at 10F-H Laws J claimed that this was 'disturbing, even nonsensical', as 'the presumption may be rebutted by proof that the child was of *normal* mental capacity for his age. If that is right, the underlying premise is that a child of *average or normal* development is in fact taken to be doli capax, but the effect of the presumption is then that a defendant under 14 is assumed to possess a *subnormal* mental capacity, and for that reason to be doli incapax. There can be no respectable justification for such a bizarre state of affairs. It means that what is by definition the exception is presumed to be the rule. It means that the law presumes nothing as regards a child between 10 and 14 except that he lacks the understanding of all his *average* peers. If that is the state of the law, we should be ashamed of it.' (Emphases added). (Note, however, that it is strikingly *abnormal* for any child aged 10-14 to be charged with a crime at all. The average charged child may lack the understanding of the average child, for these are very different categories).

[30] The Butler Report was formally titled the *United Kingdom Report of the Committee on Mentally Abnormal Offenders* (n 20). RD Mackay has proposed to update insanity's special verdict to 'not guilty on account of an aberration of *normal* mental functioning'. He continues: 'the phrase "normal mental functioning" is one which…needs no elaboration or explanation [for a jury, while] the word "aberration" can be explained…using the dictionary definition, which includes… "*deviation from normal*; mental *irregularity*…". RD Mackay, *Mental Condition Defences in the Criminal Law* (Oxford 1995) 141-142. (All emphases added).

'abnormal psychology'.[31] Abnormality, the thought continues, explains why incapacities exculpate. Call this the *normality thesis*: the thesis that incapacities exculpate (only) if and because they are abnormal.

Hart proposed something like the normality thesis. He claimed that in 'most contexts…the expression "he is responsible for his actions" is used to assert that a person has certain *normal* capacities' and that 'the possession of *the normal capacities* to conform to the requirements of law or morals [is one of…] the most prominent among the criteria of liability-responsibility.'[32] Conversely, then, the absence of these normal capacities—being abnormally incapable—might render one non-responsible, ie exculpated. Similarly, Arlie Loughnan has claimed that a defining feature of mental incapacity doctrines is their concern with mental abnormality.[33] According to Loughnan, this focus on abnormality means that the incapacity doctrines 'construct' their subjects as 'different in kind' rather than different in degree.[34] Other theorists tend reflexively to mention abnormality when discussing incapacity doctrines.[35]

There are at least three problems with the normality thesis.

First, abnormality cuts both ways. Sometimes, as with insanity, a defendants' abnormal mental functioning is both intuitively and legally exculpatory. But in other cases, abnormality has precisely the opposite effect: it seems inculpatory, or a bar to exculpation. The law frequently

---

[31] Similarly, 'Taxometric procedures provide an empirical means of determining which psychiatric disorders are typologically distinct from *normal behavioral functioning*'. Theodore Beauchaine, 'A Brief Taxometrics Primer' (2007) 36(4) J Clin Adolesc Psychol 654, 654. (Emphasis added).

[32] Hart, *Punishment and Responsibility* (n 1) 227, 265. This view echoed his and Tony Honoré's account of causation, which was '[rooted in a] conception of a human agent as being most free when he is placed in circumstances which give him a fair opportunity to exercise *normal mental and physical powers* and he does not exercise them without pressure from others'. HLA Hart and AM Honoré, *Causation in the Law* (Oxford 1959) 141. (Emphases added).

[33] Arlie Loughnan, *Manifest Madness* (Oxford 2012) 22-25. Loughnan's account is a response to Paul Robinson, *The Structure and Function of Criminal Law* (Oxford 1997) ch 5, who also explains incapacity doctrines in terms of abnormality. Robinson accounts not only for the normality of the individual but also situational normality.

[34] Loughnan, *Manifest Madness* (n 33) 22-25.

[35] RA Duff, 'Choice, Character, and Criminal Liability' (1993) 12 Law and Philos 345, 356 notes that someone terrified by a threat might be 'less able than *people of normal capacities normally are*…to assess the plausibility of the threat.' According to Susan Wolf, 'Sanity and the Metaphysics of Responsibility' in Ferdinand Schoeman (ed), *Responsibility, Character, and the Emotions: New Essays in Moral Psychology* (Cambridge 1987) 46 the law assumes 'that *normal*, fully developed adult human beings are responsible beings… [Questions about responsibility] have to do with whether a given individual falls within *the normal range*.' Joseph Raz, 'Being in the World' in *From Normativity to Responsibility* (Oxford 2011) 247 stipulatively defines disability as 'the inability to control a range of conduct that is *commonly taken to be conduct that could be expected* to be within people's domain of secure competence.' (All emphases added).

refuses to relativise its reasonableness standards to account for abnormal incapacities. Loss of control, for instance, requires a 'normal' degree of self-control/restraint.[36] Possessing (and evidencing) an abnormally low degree of self-control bars one from reliance on its exculpation. This is arguably true of the negligence standard more broadly.[37] It would be odd, to say the least, for abnormality to both bar and ground exculpation, depending on the doctrine in question. Loughnan has an answer to this worry. She excludes by definition from the scope of the incapacity doctrines any doctrine which contains a reasonableness test. Thus, she can respond that abnormality exculpates only if it doesn't form part of a reasonableness standard.[38] The problem with this reply, however, is that plenty of reasonableness standards *are* relativised to account for incapacities.[39] But even if I'm wrong about that, two further problems remain.

Second, abnormality *per se* isn't exculpatory. The simplest reason for this is that abnormality encompasses not only subnormality but also supernormality; abnormality for the worse and abnormality for the better. Someone with extremely advanced normative, epistemic, and volitional capacities cannot be exculpated on this basis. The normality thesis must only refer to subnormality. But subnormality alone isn't enough. Having some subnormal capacities isn't exculpatory without further ado. As with loss of control, having subnormal capacities of patience, diligence, or empathy is not intuitively exculpatory. If anything, the opposite. Some incapacities simply have nothing to do with culpability. Subnormal musical capacities, for

---

[36] The old law of provocation usually referred to a 'reasonable' degree of control. But this was glossed as follows in *AG for Jersey v Holley* [2005] UKPC 23 [13]: the 'powers of self-control possessed by ordinary people vary according to their age and, more doubtfully, their sex. These features are to be contrasted with *abnormalities*, that is, features not found in a person having ordinary powers of self-control. The former are relevant when identifying and applying the objective standard of self-control, the latter are not'. The normality formulation was adopted by the Coroners and Justice Act 2009 s54 (1)(c). In *Rejmanski* [2017] EWCA Crim 2061 [25]/275, Hallett LJ emphasises that 'in assessing… [this element] the defendant is to be judged against the standard of a person with a normal degree, and not an abnormal degree, of tolerance and self-restraint.'

[37] Simester et al, *Simester and Sullivan* (n 13) 162. The Australian case *McHale v Watson* (1966) 115 CLR 199, 213-214, Kitto J writes: 'The standard of care being objective, it is no answer for him, [a child,] any more than it is for an adult, to say that the harm he caused was due to his being *abnormally* slow-witted, quick-tempered, absent-minded or inexperienced. But it does not follow that he cannot rely in his defence upon a limitation upon the capacity for foresight or prudence, not as being personal to himself, but as being characteristic of humanity at his stage of development *and in that sense normal*. By doing so he appeals to a standard of ordinariness, to an objective and not a subjective standard.' (Via *Mullin v Richard* [1998] 1 WLR 1304, 1308.)

[38] This approach echoes the claims of John Gardner and Timothy Macklem that (in the context of provocation) being 'psychologically normal' means being 'apt to have their reactions judged by a rigorously objective "reasonable person" standard.' John Gardner and Timothy Macklem, 'Compassion without Respect? Nine Fallacies in *R v Smith*' [2001] Crim LR 623, 629.

[39] See ch 2 §3 for some doctrines, and ch 3 §2 for my defence of relativised reasonablenesses.

example, are neither here nor there when it comes to criminal liability. When asking why incapacities exculpate, the answer 'subnormality' cannot be the whole story. That story must refer to some other relevant incapacities. At best, subnormality might explain the requisite *degree* of incapacitation required for exculpation. That is, perhaps certain relevant incapacities exculpate only if they the individual falls below a certain capacity threshold, and perhaps subnormality might set that threshold.

Third, however, subnormality probably isn't even an appropriate threshold. That's because subnormality is a *relative* measure. Something is subnormal only in relation to the normal. For example, one has a low IQ only relative to one's position in the population: the tests are normed such that 100 represents the average test-taker. What counts as a subnormal IQ, therefore, shifts with context.[40] But the threshold below which a relevant incapacity counts as exculpatory isn't similarly relative. Neither the most capable child in a nursery nor the sanest inmate in an asylum is made (capacity-) responsible in virtue of their relative capability. Nor would be the most capable (non-human) animals left after human extinction. Incapacities exculpate because they take the incapable below an *absolute* threshold.[41]

We might yet defend the normality thesis on the basis that it picks out an appropriate threshold in, well, *normal* contexts. Legal jurisdictions are not like nurseries or asylums, after all.[42] But still

---

[40] IQ has shifted upwards over time, dubbed the Flynn effect, requiring renorming. Stuart Ritchie, *Intelligence: All That Matters* (John Murray 2015) ch 2. In the US, low IQ is a reason to stay executions, though in *Hall v Florida* 572 U.S. 701 (2014) the Supreme Court deemed a sharp cut-off (of 70) unconstitutional.

[41] A dialogue:

P: Judging when (say) normative incapacity is exculpatory presupposes that the one doing the judging is correct on the relevant normative questions. But history is full of 'moral entrepreneurs' whose beliefs seemed insane at the time but are now deemed more correct the mainstream (eg Jesus). This should give us pause before relying on a standard of normality. The law is rarely radical, but it might have better moral judgment than the *normal* person. (Leslie Green, 'Should Law Improve Morality?' (2013) 7 Criminal Law and Philosophy 473).

Q: But by what standard are we to distinguish correct contrarians from cranks? We ought to rely on a normality standard for reasons of of epistemic humility.

Verdict: A draw.

[42] Hart, *Punishment and Responsibility* (n 1) 229 notes that legal systems are 'dependent for its efficacy on the possession by a sufficient number of those whose conduct it seeks to control of the capacities of understanding and control of conduct which constitute capacity-responsibility. For if a large proportion of those concerned could not understand what the law required them to do or could not form and keep a decision to obey, no legal system could come into existence or continue to exist.' (Though we could quibble this: after all, a 'large proportion' of the world are children: globally, 25% are under 14, rising towards 50% in some countries. CIA, *CIA World Factbook* <https://www.cia.gov/library/publications/the-world-factbook/geos/xx.html>.)

a question remains: *how* subnormal must one's capacities be to exculpate? It cannot just be below the median: that would be far too broad. Nor would picking out a low percentage. Some exculpatory incapacities are very rare (eg Korsakoff's syndrome), while others are quite common (eg depression).[43] Once again, picking out a relative measure by which to judge exculpation seems inappropriate, even as a threshold test.

But perhaps I am misinterpreting the normality thesis. There is experimental evidence that judgments of 'normality' don't simply describe a statistical average. Rather, judgements of normality combine descriptive statistical averages with prescriptive ideals.[44] Thus, the normality thesis can be interpreted to say that incapacities exculpate if they fall below some hybrid statistical/normative threshold.[45] Granting that this is correct, however, we still need a separate account to explain what counts as prescriptively ideal. As a result, the normality thesis becomes very thin indeed. It doesn't explain why incapacities exculpate. Rather, in attempting to explain *when* an incapacity (specified elsewhere) falls below an exculpatory threshold, the normality thesis relies heavily on an account (specified elsewhere) of the ideal capacity in that otherwise specified domain. Even at its highest, then, the normality thesis doesn't explain much. Perhaps we should pay some attention to statistically average capacities when deciding why incapacities exculpate. But normality can be but a small part of the story.[46]

I don't think that the normality thesis succeeds as a general explanation for why incapacities exculpate. But something like the normality thesis might succeed if localised to a particular variety of exculpatory incapacity. The idea is this. Only some of our bodily movements (and thoughts, etc) count as actions. Plausibly, a necessary condition for a movement to count as an action is that it occurs via the brain's dedicated action selection system through an appropriate causal pathway. You moving my arm does not an action make, for it doesn't derive from an appropriate causal pathway. But not all internal causes count as actions: spasms and

[43] Technically, per the last chapter, I mean 'the exculpatory incapacity derived from those aetiologies' – but even thus substituted some incapacities will be much more common than others.

[44] Adam Bear and Joshua Knobe, 'Normality: Part Descriptive, Part Prescriptive' (2017) 167 Cognition 25. This account has been applied to the legal concept of reasonableness in Kevin Tobia, 'How People Judge What is Reasonable' (2018) 70 Ala L Rev 293.

[45] Sometimes criminality and criminals are *defined* as (mentally) abnormal. But, as Glanville Williams pointed out, this implies an (implausibly) broad concept of normality. He distinguished exculpatory abnormality ('psychotics, neurotics' etc.) from 'socially normal' criminals. The latter would be fully responsible. Glanville Williams, 'The Definition of Crime' [1955] Current Legal Problems 107, 113.

[46] A further practical objection is that 'The distinction between what is normal and abnormal is one of degree and can be difficult to draw.' Law Commission, *Partial Defences to Murder* (Final Report, 2004) [5.74] (regarding diminished responsibility).

'actions' derived from basal ganglia misfires are not plausibly genuine actions. But distinguishing such movements from genuine actions is very tricky. What counts as an 'appropriate' causal pathway is hard to pin down. Action theory isn't easy. But one plausible view is that the action must be *controllable*; that is, in control of our executive faculties. Neurologically, this seems to mean that the inputs to the basal ganglia derive from the (pre-) motor cortex.[47] In other words: from the *normal* causal pathway. Thus: sometimes normality explains what counts as an action in the first place. Abnormal causal pathways (basal ganglia misfires etc) mean no *action* is performed. That a movement is a non-action often explains why it exculpates. And the legal route to this exculpation sometimes occurs via the incapacity rules rather than a mere absence of an actus reus, where the movement (falsely) *appears* to be a genuine action.[48] This is far from a complete explanation of why incapacities exculpate. But it at least salvages some truth in the normality thesis.

3 Response-ability

So much for the normality thesis as a unifying explanation for why incapacities exculpate. Insofar as normality is relevant, it is only to specify when a potentially exculpatory incapacity is sufficiently severe as to exculpate. We still need to ask in virtue of what an incapacity is potentially exculpatory. The most parsimonious explanation would refer to a master exculpatory incapacity. The two options I'll consider are Response-ability and Conformability.

The root sense of 'responsibility', etymologically, refers to an ability *to respond*.[49] This etymological root, some claim, reflects the philosophical truth that Response-ability is the central or basic sense of responsibility.[50] It might follow that an absence of Response-ability is the core exculpatory incapacity.

---

[47] For a philosophically oriented discussion of the empirics here, see Timothy Schroeder, Adina Roskies, and Shaun Nichols, 'Moral Motivation' in John Doris (ed), *The Moral Psychology Handbook* (Oxford 2010) §3.

[48] This also explains the (narrow) truth of the claim, dismissed in the last chapter, that some aetiologies morally transform incapacities to render them exculpatory. Sometimes the peculiar causal pathway explaining our movements—their aetiology—really does make it the case that those movements are non-inculpatory—even if, from the outside, there is no obvious basis of exculpation. But even here there is an explanatory intermediary, viz the absence of any action. See text at ch 4 fn 83.

[49] Hart, *Punishment and Responsibility* (n 1) 265.

[50] Gardner refers to the idea as 'basic responsibility': John Gardner, 'The Mark of Responsibility' in *Offences and Defences* (Oxford 2007) 182.

But to what must one respond, and how? Pigeons can respond to pigeon calls with an appropriate response, but that hardly renders them responsible in the relevant sense. As John Gardner puts it, Response-ability requires the ability to respond *to reasons*, *by offering justifications and excuses* for our conduct.[51] He claims that basic responsibility (or Response-ability)

> It is an ability which straddles the temporal gap between the wrong or mistake and the trial or recrimination, and which also straddles the conceptual gap between the ability to respond to reasons in what one originally does or thinks or feels, etc., and the ability to use those same reasons in explaining what one did or thought or felt.[52]

I noted in chapter 3 that Gardner's insistence on compounding two distinct abilities/capacities when formulating basic responsibility can lead to confusion.[53] Whether one is able to respond to accusations *ex post* (being fit to plead) is quite distinct from being able to respond to reasons *ex tempore*, at the time of action (being sane). But even if Gardner were right to compound the two qua 'basic responsibility', it would be wrong to do so when considering which incapacities exculpate. For clearly the relevant capacity is the capacity to respond to reasons *at the time of action*. My murdering another is no less culpable for the fact that I *later* become unfit to plead. Having said that, being able simply 'to respond to reasons' isn't enough. For it doesn't tell us *how* one must respond to reasons. Pigeons respond to their reason to eat by eating. That doesn't make them Response-able. Hence Gardner's claim that Response-ability 'is none other than an ability to offer justifications and excuses.'[54] As we will see, however, this remains unclear in a way that undermines the plausibility of the view.

Many individuals are able to respond to reasons despite intuitively having exculpatory incapacities. Schizophrenics, for example, often confabulate intricate narratives (and conspiracies) to explain delusional beliefs, eg that people on the news are talking about them. Likewise, they might attempt to justify erratic movements in terms of dodging phantom attackers.[55] They can offer justifications and excuses. If such an individual committed an offence to escape their (fictitious) tormenters, they might be able to offer a perfectly coherent,

---

[51] Hart, *Punishment and Responsibility* (n 1) 265.

[52] Gardner, 'The Mark of Responsibility' (n 50) 184.

[53] Ch 3 fn 8.

[54] Gardner, 'The Mark of Responsibility' (n 50) 182.

[55] Two recent first-personal accounts of delusional confabulation are Susannah Cahalan, *Brain on Fire: My Month of Madness* (Simon and Schuster 2012) and Esmé Weijun Wang, *The Collected Schizophrenias* (Graywolf 2019).

even plausible, justification or excuse for their conduct. Yet in many cases, they would, intuitively, be exculpated on the basis of incapacity.

The Response-ability proponent could deflect these counterexamples by claiming that wrongdoing in the grip of delusions only exculpates where the wrongdoing would have been justified or excused had the delusions been true. They are not exculpated on the basis of incapacity. Rather, they are exculpated on the basis of a non-incapacity-based defence—self-defence, duress, necessity, etc—which they mistakenly believed applied. We don't seek incapacity-based exculpation for mistaken self-defence attributable to mere perceptual error. Nor, then, does it count as incapacity-based exculpation just because that perceptual error derived from an incapacity. But this response doesn't account for (counterfactual) relativisations. If the very reason for relativising or omitting a standard (of belief) is to account for incapacities, then delusional mistaken self-defence should qualify as incapacity-based exculpation. Then we have an exculpatory incapacity despite no lack of Response-ability. Thus, Response-ability can't be the master incapacity which explains why incapacities exculpate.

My account of (counterfactual) relativisations is controversial. Someone like John Gardner could simply reject my characterisation of those doctrines. But other counterexamples don't leave this get-out. In some cases, individuals confabulate explanations—justifications and excuses—unrelated to delusional perceptions. The starkest examples arise from neurological disorders such as agnosia or Korsakoff's syndrome. As Oliver Sacks recounts in his vivid case studies, individuals unaware of their circumstances, eg due to memory loss or failure to identify faces, might yet provide comprehensive but utterly incredible explanations as to their circumstances or conduct.[56] They surely ought to be exculpated if they committed some crime in virtue of their bizarre beliefs (eg if they unreasonably believed the person they sexually touched was their spouse). But that exculpation doesn't derive from a perceptual error, strictly speaking.

Gardner might resist this sort of example for a different reason. As he puts it, 'Basically responsible agents can't always give a rationally *acceptable* account of themselves, but they can always give a rationally *intelligible* one.'[57] These incredible confabulators might offer justifications and excuses, but not intelligible ones. That, Gardner can claim, is why they are

[56] Oliver Sacks, *The Man Who Mistook His Wife for a Hat* (Summit 1985).

[57] John Gardner, 'Hart and Feinberg on Responsibility' in Matthew Kramer et al (eds), *The Legacy of HLA Hart* (Oxford 2008) 123. (Emphasis in original.)

to be exculpated. However, plenty of confabulated explanations are intelligible. Indeed, confabulators (especially those without hallucinatory delusions) often offer the most plausible hypotheses they can find.[58] Alas, those hypotheses happen to be false. They lack a proper connection to reality.

Here Gardner's account of basic responsibility, of Response-ability, becomes trickier. For he denies that (some) confabulators of plausible explanations are basically responsible. Gardner claims that

> An ability to offer justifications and excuses…implies an ability to *have* a justification or excuse… [B]asic responsibility is an ability to give a rational explanation for one's actions without giving one's actions any rational explanation that they didn't actually have, ie without inventing reasons for what one did.[59]

In a later paper, he notes that Response-ability is relevant to us in our 'role as a rational agent, an agent subject to reasons.'[60] But for confabulators and delusional individuals

> There are no facts they can point to either as reasons for doing as they did or as reasons for their being disposed to do it. There are only imaged facts… [T]he only "reasons" they can invoke—and they are reasons only in a peripheral sense—are not facts but figments of their fevered imaginations.[61]

In other words, Gardner adds a *truth proviso* to the definition of Response-ability given above. Response-ability is not merely the ability to offer *some* justifications and excuses. It is, rather, the ability to offer *existent* reasons for our conduct; to offer *true* justifications or excuses. In that case the plausible confabulator will lack Response-ability, and thus will be exculpated. Thus: the view does not suffer from the harshness I suggested.

But Gardner seems to disavow the truth proviso in the very same paper:

---

[58] Sacks, *The Man Who Mistook His Wife for a Hat* (n 56) ch 12 mentions the case of a grocer who would fail to identify him (Sacks), but would offer various attempted identifications according to the limited evidence available: his prior indicated a customer, the white coat evidenced a butcher, his appearance (presumably) implied a mechanic dressed as a doctor, or perhaps he was a *different* doctor...

[59] Gardner, 'The Mark of Responsibility' (n 50) 184

[60] Gardner, 'Hart and Feinberg' (n 57) 124.

[61] Gardner, 'Hart and Feinberg' (n 57) 125-126.

> [Basic] responsibility…does not lie in our ability to provide *successful* justifications and excuses, or even *credible* justifications and excuses. It lies in our ability to provide justifications and excuses full stop.[62]

If I understand these passages correctly, Gardner seems both to endorse and to deny the claim that Response-ability entails the ability to offer true justifications or excuses.[63] In other words, his account of Response-ability equivocates between being (1) the capacity to respond to *actual* reasons (with the truth proviso) and (2) the capacity to respond to *merely apparent* reasons (without the truth proviso).

This equivocation is for good reason: we should reject both versions. If we reject the truth proviso, then we say that one is Response-able if (and only if) one can respond to merely apparent reasons. But in that case the worry about plausible confabulators remains outstanding. They offer apparent reasons, but their disconnect from reality surely renders them non-culpable. Response-ability would be too harsh a criterion of exculpation.

By contrast, if we accept the truth proviso, then we say that one is Response-able (if and) only if one can respond to actual reasons. But this is an implausibly restrictive condition as to who counts as Response-able. All of us offer false explanations for our conduct all the time. Not because we lie, in most cases. Rather, because we are often ignorant as to the actual drivers of our emotions and behaviour. We claim to be angry, when in fact we're scared. We eat, thinking we're hungry, when in fact we're thirsty. We explain compulsive nervous hair-touching by a fictitious need to move hair away from our face. These everyday confabulations are well-known both to psychology and philosophy.[64] In Gardner's words, we give our actions rational

---

[62] Gardner, 'Hart and Feinberg' (n 57) 123.

[63] Gardner could perhaps reconcile these passages be disambiguating the temporal frame and confirming that an ability need not be exercised. That is, D might *have been* able to have a justification (be justified) even if, as it transpired, he did not act justifiably. He can be response-able in virtue of being able to offer the justification which he was able to have, though in fact did not. This sits somewhat awkwardly with Gardner's views about unexercised capacities discussed later. At any rate, I reject both interpretations.

[64] Psychologists label this 'affect misattribution'. In one famous study, participants on a 'fear-arousing suspension bridge' evidenced greater attraction to an experimental confederate than those on a duller bridge, a result interpreted to mean that they misattributed their physiological fear symptoms as sexual arousal. Donald Dutton and Arthur Aron, 'Some Evidence for Heightened Sexual Attraction Under Conditions of High Anxiety' (1974) 30 J Pers Soc Psych 510. For a modern survey and attempted explanation of such processes: Keith Payne et al, 'A Process Model of Affect Misattribution' (2010) 36 Pers Soc Psych Bull 1397-1408. (I rely on the familiarity of everyday examples in preference to the precise claims of the psychology literature as the latter might not replicate.) As for philosophers, Nietzsche famously claimed that our motives are unknown to ourselves: Friedrich Nietzsche, *Daybreak: Thoughts on the Prejudices of Morality* (1881, Clark and Leiter eds, Cambridge 1997) §116; Brian Leiter,

explanations that they didn't actually have. If Response-ability is interpreted with the truth proviso, then all of this everyday conduct would be deemed non-Response-able and hence non-culpable. That includes all the many crimes committed in the grip of self-delusional rationalisation. If we accept the truth proviso, Response-ability would be too lenient a criterion of exculpation.

The Response-ability story thus faces a dilemma with no good answer. If it rejects the truth proviso then it would be implausibly harsh. It would fail to exculpate those whose explanations have lost touch with reality. But if it accepts the truth proviso then it would be implausibly lenient. It would exculpate anyone who acted for non-existent reasons. Given this, Response-ability seems like a poor explanation for why incapacities exculpate.

4 Conformability

What of the alternative, then, the Conformability story? The idea is familiar enough. If someone is incapable of conforming to some norm, we tend not to consider them culpable for failure to conform. As Hart famously put it,

> those whom we punish should have had, when they acted, the normal capacities, physical and mental, *for doing what the law requires and abstaining from what it forbids*, and a fair opportunity to exercise these capacities. Where these capacities and opportunities are absent, as they are in different ways in the varied cases of accident, mistake, paralysis, reflex action, coercion, insanity, etc, the moral protest is that it is morally wrong to punish because "*he could not have helped it*" or "*he could not have done otherwise*"[65]

This position is often repeated by legal authorities.[66] We're not culpable for arriving late if all transport options are suddenly shut down without warning. Why? Because we couldn't make it on time. Doctors are not culpable for failing to save patients if nothing could have been done. Admittedly, this intuition doesn't always hold. If the doctor couldn't save the patient as

Nietzsche on Morality (Routledge 2002) 101-104; Brian Leiter, *Moral Psychology with Nietzsche* (Oxford 2019) chs 5-7.

[65] Hart, *Punishment and Responsibility* (n 1) 15 (emphases added).

[66] ALI, *Model Penal Code* (1962) 4.01 'A person is not responsible for criminal conduct if at the time of such conduct as a result of mental disease or defect he lacks substantial capacity…to conform his conduct to the requirements of law.' Hart's quote was endorsed in *Loake v CPS* [2017] EWHC 2855 (Admin) [35]. The Law Commission, *Criminal Liability: Insanity and Automatism* (Discussion Paper, 2013) [1.52] refers to the capacity 'to obey the law'. I prefer the 'conform' formulation as obedience is usually thought to require conformity *for the sake of the law*, which is less plausible as a condition for exculpation.

a result of his own voluntary intoxication, then we would rightly deem him culpable. But the explanation for this is simply that he is separately culpable for disabling his capacity.[67] Conformability entails that D is not culpable for [not] Φing if D cannot [but] Φ.

How does Response-ability differ from Conformability? If we want to explain why wrongdoing performed in the grip of psychosis ought to be exculpated, we could either say that the psychotic individual couldn't properly respond to reasons, or we could say that the individual couldn't conform to the law. In many cases, perhaps most, the two explanations offer the same answer. But not all. If A pushes B into C, B might be able to respond to reasons ('I don't want to hurt you, look out!') without being able to conform to the law's requirements (simplified: to avoid contact).[68] Response-ability would (apparently) deny exculpation that Conformability offered.

Having said this, John Gardner (at first) agrees that B, who 'might well have been a fridge…is not basically responsible for having knocked C over.'[69] The natural explanation for why B is not culpable for his contact with C is that B couldn't help it. That is the explanation offered by the Conformability story. But Gardner, contra Hart, calls this 'misleading':

> [B] *is*…basically responsible for having knocked C over [if]… there was some action Φ-ing such that, if only B had Φ-ed in the course of his falling, he would have averted C's being knocked over by him… [It] is not necessary for Φ-ing to be an action which V had either the capacity or the opportunity to perform. All that is necessary is that if B had Φ-ed (even *per impossibile*) C would not have been knocked over.][70]

The reason:

> If there was such an action—such as pushing C out of the way, or shouting "watch out C!", or blocking his own fall—then B was still, in (not averting) his knocking C

[67] See the discussion of prior fault in ch 4 §3.2. Would the doctor be culpable if, but for the intoxication, he still would not have been capable of saving the patient? He wouldn't be criminally liable for that failure in English criminal law, for want of causation. But he might yet be culpable, and criminally liable under a different heading, for the very act of becoming intoxicated.

[68] The example is from Gardner, 'Hart and Feinberg' (n 57) 124. I discuss Gardner's discussion of the case below.

[69] Gardner, 'Hart and Feinberg' (n 57) 124.

[70] Gardner, 'Hart and Feinberg' (n 57) 124; section in square brackets from fn 5. Gardner phrases this condition (double) negatively, but he must endorse the positive formulation, as evidenced by the next quote. (First emphasis added).

> over, occupying his role as a rational agent, an agent subject to reasons… We can ask [him] for a justification or an excuse. In asking for such a justification or excuse, we are treating B as basically responsible for (not averting) his having knocked C over. His (not averting) his having knocked C over was a manifestation of his rational agency even if his falling in the first place was not.[71]

Without any requirement that Φing is possible to perform, there is always, necessarily, some action that, if performed, would avert any given outcome. For there is always some action open to a superman or a deity or so on that would avert any and all bad outcomes.[72] But it seems odd, to say the least, to think that this renders mere mortals like B basically responsible for their conduct in circumstances where no human could have averted knocking into C. So perhaps Gardner meant something more restricted, something like 'some *humanly possible* action' or 'some *plausibly performable* action' or the law's favoured 'some *reasonable* action'.[73] But even if so, it's hard to see how that abstract possibility bears on the basic responsibility of an agent for whom those actions are entirely impossible. Imagine that B's knocking into C could have been averted by throwing out his arms and catching his body. Now imagine that B was too elderly to react fast enough, or too physically weak to catch himself, or even that he didn't have any arms. Is B to be considered basically responsible for his failure?[74] Gardner says yes despite the fact that B features in the collision no more agentially than a fridge. That is unintuitive, to put it mildly.[75]

---

[71] Gardner, 'Hart and Feinberg' (n 57) 124.

[72] Cases of nomological or logical impossibility excepted, perhaps.

[73] Elsewhere Gardner says reasons for action must be 'possible in principle for people to act on them', perhaps implying humanly possible. But he emphasises that 'the words "in principle" [should be] taken seriously', and switches terminology from 'people' to 'rational being', perhaps including supermen and the like. John Gardner, 'The Wrongdoing that Gets Results' (2004) 18 Philosophical Perspectives 53, 54, as picked up by RA Duff, 'Cliff-top Predicaments and Morally Blemished Lives' (2019) 19 JRLS 125, 126 and 130ff.

[74] Gardner emphasises that in these circumstances 'We begin by asking: 'Why did you do that?"; and indeed, that this question is right and proper to ask. But the intuitive appropriateness of this reaction is entirely context specific. If gangster A cruelly hurls frail pensioner B into C, and C is fully aware of what transpired, it would be shockingly inappropriate for C to demand an explanation of B. Gardner admits that it might be inapt to demand an explanation of the basically responsible for supervening reasons (eg lack of standing): Gardner, 'Hart and Feinberg' (n 57) 125 fn 8. But that response fails to deflect the fundamental objection that B is simply not basically responsible; that B need not offer any explanation for his crashing into C, prima facie or otherwise. This point echoes Luis Duarte d'Almeida, 'O Call Me Not to Justify the Wrong': Criminal Answerability and the Offence-Defence Distinction' (2012) 6 Crim Law and Philos 227.

[75] Gardner would not be impressed by my appeal to a widespread intuition of plausibility; he claims to count this as a mark against the truth of a philosophical proposition. John Gardner, 'As Inconclusive as

But perhaps I am talking across Gardner. Gardner's claim is that B remains *basically responsible*. He isn't saying that B is *culpable*. B could be basically responsible and yet not be culpable: if he could *excuse* his crashing into C. According to Gardner, to offer an excuse is precisely to *assert*, not to deny, basic responsibility.[76] Thus: Gardner didn't mean to deny that B is non-culpable, and thus his view isn't as implausible as I have suggested. But if this were Gardner's reply, he would still need to explain why B is exculpated. What exactly is B's excuse? B doesn't have any rational defence like self-defence or duress. Legally speaking, B would be considered a mere tool of A. His crashing into C would not be deemed an 'act' (or a 'voluntary act') and thus would not fulfil the actus reus required of most offences. But to deny that one acted at all is precisely to deny basic responsibility in Gardner's sense. It is no answer to disambiguate, to say that B is not basically responsible *for falling* but is basically responsible *for knocking over C*. For we still need to know what grounds B's exculpation *for knocking over C*.[77] Gardner has no answer in terms of basic responsibility whenever there was the possibility—even if not for B—of performing an averting action. Still: can't Gardner have B offer a different excuse, one that doesn't implicate basic responsibility? Perhaps. But Gardner insists that all excuses are rational explanations; that they do not rely on claims of incapacitation.[78] And that means he cannot explain this most ordinary case of incapacity-based exculpation.[79]

By contrast, nothing could be easier to explain for the Conformability story. B could not avoid crashing into C, and that incapacity—barring any prior fault (like provoking A)—grounds his

Ever' (2019) 19 JRLS 204, 209. Leaving aside any doubts about this iconoclasm, Gardner's argument appeals to nothing but the plausibility of the aptness of asking B for an explanation. That it is not, in fact, plausible must count against the view.

[76] Gardner, 'The Mark of Responsibility' (n 50) 181-182.

[77] B might be responsible *for a failure to feel certain emotions*; even, perhaps, up to agent-regret. But it doesn't follow that B is responsible for the (non-)action that precipitated the responsibility to feel such emotions. Duff makes a related point in terms of reasons in Duff, 'Cliff-top Predicaments' (n 73) 132.

[78] I discuss Gardner's normative scepticism about incapacity-based excuses in the next section.

[79] Put another way, Gardner claims that one is blameworthy if four conditions are present: (1) wrongdoing, (2) basic responsibility, (3) no-justification, (4) no-excuse. In the case under discussion, Gardner claims that B is basically responsible, and lacks justification and excuse. Gardner also considers wrongdoing to be strict. Thus: B must be blameworthy on Gardner's schema. That is the wrong result. I've claimed that the error lies with criterion (2): in fact, B is not basically responsible, as he couldn't conform to the law. Another response, however, might be to query criterion (1), ie whether wrongdoing ought to be considered strict. But that is a discussion for another day.

exculpation. This explanation is simple and intuitive. It's endorsed by various theorists.[80] It is the best explanation available as to why incapacities exculpate.

With this explanation in hand, we can clarify and resolve some difficulties with the three-incapacities story presented above.

For one thing, what are we to make of the fact that different formulations of those incapacities, often with widely differing implications, are used by different authors? To say that normative incapacity requires a failure fully to appreciate moral wrongfulness is not at all the same as saying it requires the absence of an abstract awareness of legal or social condemnation. Can those two statements really refer to one underlying incapacity, normative incapacity? That may depend on the authors in question. But there is a way to understand these claims compatibly. We can take them to be disagreeing about the extent to which their rival descriptions of normative incapacity imply an absence of Conformability regarding a particular norm. Perhaps a full appreciation of moral wrongfulness is required for conformity. Or perhaps not: perhaps conformity remains possible until and unless the agent cannot grasp even social disapproval.[81] On this interpretation of the dialectic, these rival views as to the content of the 'normative incapacity' label are really disagreeing about the component parts of a shared, more fundamental, Conformability story.[82]

In other words, the three incapacities mentioned above are not incapacities of fundamental relevance to exculpation. Instead, they pick out fine-grained incapacity descriptions which tend to imply an absence of the fundamentally relevant capacity: Conformability. Those incapacities are exculpatory only in general; only to the extent that they imply an absence of

[80] For recent conclusions contra Gardner's argument: Duff, 'Cliff-top Predicaments' (n 73); Michelle Madden Dempsey, 'What We Have Reason to Do: Another View from the Cliff-Top' (2019) 19 JRLS 141.

[81] Similar to the controversy as to what normative incapacity entails, there is (some) controversy as to whether normative incapacity exculpates at all. Elizabeth Harman, for example, claims that ancient Greek slaveholders are likely not exculpated, despite lacking the ability to appreciate the immorality with their (commonplace) slaveholding, given their society. (One is not exculpated, Harman thinks, if one demonstrates 'de re insufficient care about what is morally significant'). Elizabeth Harman, 'Does Moral Ignorance Exculpate?' (2011) 24 Ratio 443, 460. Others, by contrast, hold that this moral ignorance would exculpate, just as factual ignorance would (eg ignorance that one's actions in fact caused slavery): Gideon Rosen, 'Culpability and Ignorance' (2003) 103 PAS 61, 64 (calling this the 'parity thesis'). Others go further, intuiting that a modern-day dictator's son insulated from criticism might be exculpated for terrible dictatorial decisions: Susan Wolf, 'Sanity and the Metaphysics of Responsibility' (n 35) 53-56. One way to understand these conflicting intuitions is that the authors disagree about the extent to which fine-grained incapacity descriptions correspond to a deeper exculpatory incapacity.

[82] I'm not claiming that this is how the disputants conceptualise their own arguments.

Conformability. They are of only indirect, heuristic relevance. This explanation also allows us to make sense of Hart. Above I quoted Hart twice: both for his discussion of the three incapacities and for his discussion of the Conformability story. We might ask of Hart: which is exculpatory? The three incapacities or Conformability?[83] But this is a false choice. Hart advocates both. The three incapacities *usually* exculpate, but they do so *because* they imply an absence of Conformability.

5 Ought implies can[84]

Conformability is intuitively attractive in explaining why (and which) incapacities exculpate. But this very intuitiveness means that it has received plenty of attention, including critical attention, from philosophers. This attention has focused on two areas in particular: the 'ought implies can' principle and free will.

Recall

> *Conformability*: D is not culpable for [not] Φing if D cannot [but] Φ

This can be reached as a conclusion from a common version of the 'ought implies can' principle (OIC) combined with a supplemental premise. Start with OIC:

> *OIC:* D ought to Φ only if D can Φ

OIC's contrapositive:

1) If D cannot Φ then not (D ought to Φ)[85]

---

[83] And, for Hart, 'fair opportunity'.

[84] This section benefitted from two unpublished manuscripts by Frederick Wilmot-Smith.

[85] Modus tollens can fail if multiple senses of 'can' are used. This is brought out more clearly with a version of OIC where the implication refers to a conditional:

> OIC*: If D ought to Φ then D cannot lack the [normative/transcendental] ability to Φ
>
> 1) If D lacks the [physical] ability to Φ then not (D ought to Φ)

The different senses of 'can' between the two statements mean they are not equivalent. Kant arguably endorsed something like OIC*: John Gardner, 'Reasons and Abilities: Some Preliminaries' (2013) 58 Am. J. Juris 74. By contrast, others took OIC to mean (1): GE Moore, *Philosophical Studies* (Harcourt, Brace & Co 1922) 317-319 (via Greg Caruso, 'Skepticism About Moral Responsibility' in Edward Zalta ed, *The Stanford Encyclopedia of Philosophy* (Spring 2018 edn) §3.3 and Peter Vranas, 'I Ought, Therefore I Can' (2007) 136 Phil Stud 167, 183). I return to the (alleged) Kantian version below.

The supplemental premise links 'oughts' to culpability:

S: D can be culpable for Φing only if D ought not to Φ[86]

The inverted contrapositive:

2) If not (D ought to Φ) then D cannot be culpable for not Φing

From these premises, an argument for Conformability:

1) If D cannot Φ then not (D ought to Φ)
2) If not (D ought to Φ) then D cannot be culpable for not Φing
3) ∴ If D cannot Φ then D cannot be culpable for not Φing
4) ∴ (By inversion) D is not culpable for not Φing if D cannot Φ (*Conformability*)[87]

Those who accept both (this version of) OIC and S ought also to endorse Conformability.

What of those who reject (this version of) OIC? Do they *reject* Conformability? That doesn't follow. We might not be culpable for not Φing, due to inability, despite the fact that we ought to Φ. Inability might deflect culpability by *excusing* rather than by denying wrongdoing.[88] Indeed, I find this view more plausible than the argument given above. My point was simply that *if* one accepts OIC and S, as many philosophers do, then one ought also to endorse Conformability.

---

[86] This claim is endorsed by Stephen Darwall, *The Second-Person Standpoint* (Harvard 2006) 27-28 (A 'connection to accountability is part of the very concept of moral obligation… [It] make[s] no sense to blame someone for doing something [they had] sufficient reason to do'. Darwall cites various philosophers, including Mill, in agreement. (Reference via Frederick Wilmot-Smith). It is rejected by Peter Vranas, 'I Ought, Therefore I Can' (n 84) 192-194. Typical counterexamples involve cases where D takes an unjustifiable risk that luckily turns out well. The intuition is that D is blameworthy but did not act wrongly. The plausibility of such counterexamples turns on how we individuate D's action/outcome and how we understand wrongs. I won't take sides on those issues.

[87] I take it as analytic that 'cannot be X' entails 'is not X'. To spell out the inference from (3) to (4):

3) If D cannot Φ then D cannot be culpable for not Φing

3A) If D cannot be culpable for not Φing then D is not culpable for not Φing

3B) ∴ If D cannot Φ then D is not culpable for not Φing

4) ∴ (By inversion) D is not culpable for not Φing if D cannot Φ

[88] Some claim that OIC is better interpreted purely in relation to blaming judgments: Walter Sinnott-Armstrong, "Ought' conversationally implies 'can" (1984) 93 Phil Rev 249; Vladimir Chituc et al, 'Blame, not ability, impacts moral 'ought' judgments for impossible actions: toward an empirical refutation of 'ought' implies 'can" (2016) 150 Cognition 20-25.

Thus far I have been referring to the literature's dominant version of OIC. But John Gardner claims that OIC's most famous proponent—Kant—endorsed the opposite thesis:

OIC*: If D ought to Φ then D *cannot lack* the ability to Φ[89]

While OIC moves from facts about our abilities to a conclusion about what we ought (not) to do, OIC* moves from facts about what we ought (not) to do to a conclusion about our abilities. These theses would amount to the same thing if OIC*'s consequent is read to restrict its antecedent, such that

i. If D ought to Φ then D *cannot lack* the ability to Φ (*OIC**)
ii. D lacks the ability to Φ
iii. ∴ Not (D ought to Φ)
iv. ∴ If D cannot Φ then not (D ought to Φ) (*OIC*)[90]

If Gardner's Kant meant something different, he must mean that the ordering of the premises suggests a grounding relation, such that OIC* amounts to:

OIC^{G}: That we ought to Φ *makes it the case* that we can Φ.

If OIC^{G} is correct, Conformability is untenable. It would get the story precisely backwards. Abilities do not limit what we ought to do. Rather, what we ought to do sets the limits of our abilities. But OIC^{G} is implausible. That I ought to save a drowning child tells us nothing about what I am physically able to do. No amount of obligatoriness changes the fact that I cannot swim. OIC^{G} is only remotely plausible if 'ability' is given not a physical but a normative or transcendental meaning. At any rate, it's not obvious that Kant, or anyone else, actually endorses OIC^{G} as opposed to OIC*/OIC.[91]

---

[89] Gardner, 'Reasons and Abilities' (n 85).

[90] Assuming, again, the inference in fn 87.

[91] My aim is not exegetical, but a brief note on Kant. He never used the canonical 'ought implies can' formula, and was never particularly clear about his thesis. But I interpret him to mean

KOIC: The moral law commands us to Φ only if we can Φ

This is close to the traditional OIC reading. The appearance that Kant endorses OIC* comes from neglecting that 'commands *us*' is to be distinguished from what the moral law requires of *other* rational agents. On that question, Kant seems to endorse something like:

KOIC*: If the moral law requires Φ then not *all* rational agents can *lack* the ability to Φ.

But this does not contradict (K)OIC. In other passages Kant emphatically rejects 'excuses' for moral failure which cite common human foibles or psychological resistance. But I read such passages as (empirically) denying that such excuses amount to an incapacity to conform, rather than denying the relevance of real incapacities to conform. For the (exhaustive list of relevant) passages on which my

OIC furnishes Conformability with some allies and, as far as I can tell, no foes.

6 Free will[92]

According to Conformability, we are not culpable for [not] Φing if we cannot [but] Φ. This raises obvious philosophical concerns about free will. For Conformability is, in essence, the *principle of alternate possibilities* (PAP), according to which we can only be culpable if we could have done otherwise than we actually did.[93] Many compatibilist philosophers reject PAP, and thus also Conformability.

(A note of caution before we begin: in this subsection, I'll consider how various arguments about free will interact with Conformability. My presentation of these arguments won't consider the (often detailed) argumentative support for their premises. They are not intended to be the best nor the most comprehensive formulations of these arguments. They are only intended to show how Conformability fits into the landscape.)

I'll consider the rejections of PAP in a moment. What, though, of those who accept something like PAP? They tend to be incompatibilists. According to *Hard Determinism*:

5) We are not culpable for Φing if we cannot but Φ (*Conformability*)
6) For all Φ, we cannot but Φ[94] (*Determinism*)
7) ∴ For all Φ, we are not culpable for Φing (*Hard Determinism*)

reading is based, see Robert Stern, 'Does 'Ought' Imply 'Can'? And Did Kant Think It Does?' (2004) 16 Utilitas 42, 52-61 (albeit Stern does not quite endorse my reading).

[92] The free will literature refers variously to scepticism about free will, free action, moral responsibility, and blameworthiness. I'll interpret these concepts as either synonyms for or necessary conditions of culpability.

[93] Harry Frankfurt, 'Alternate Possibilities and Moral Responsibility' (1969) 66 The Journal of Philosophy 829, 829.

[94] As many have pointed out, determinism is at best a convenient shorthand. The current empirical picture is that determinism is false. Its truth is at best an open question, dependent (it seems) on whether apparent quantum indeterminism (as described by Bell's inequality) ought to be interpreted to infer a deterministic set of multiple worlds (the Many Worlds, or Everett interpretation of quantum mechanism). These waters are too deep for me. The real problem is that of *causal mechanism* (other than agent causation). That my Φing is stochastic or random is cold comfort: Daniel Dennett, 'Mechanism and Responsibility' in Gary Watson (ed), *Free Will* (Oxford 1982) ch X. I refer to 'determinism' as a shorthand.

If Determinism is true, then it is always the case that we cannot but do what we in fact do. Conformability then entails that we are never culpable.[95] My culpability principle then entails that we should object to all convictions. This is an uncomfortable conclusion. It prompts many philosophers to reject Conformability/PAP.[96] Uncomfortable, yes, but no contradiction of Conformability. *Libertarians* avoid this uncomfortable conclusion by denying Determinism:

5)	We are not culpable for Φing if we cannot but Φ	(*Conformability*)
8)	Sometimes we can but Φ [ie we can (Φ or not Φ)]	(*Anti-determinism*)
9)	∴ Sometimes we are culpable for Φing	(*Free will*)

If Libertarianism is correct, then obviously determinism presents no problem for Conformability, for determinism is false. I have qualms about both views, but neither undermine Conformability.[97]

What then of those who reject PAP? According to some *Compatibilists*:

6)	[Arguendo] For all Φ, we cannot but Φ	(*Determinism*)
9)	Sometimes we are culpable for Φing	(*Free will*)
10)	∴ Sometimes we are culpable for Φing but cannot but Φ	(*Compatibilism*)
11)	∴ Not (We are not culpable for Φing if we cannot but Φ)	(*Anti-conformability*)

[95] A conclusion embraced, among others, by Galen Strawson, 'The Impossibility of Moral Responsibility' (1994) 75 Phil Stud 5.

[96] Hard Determinism entails that we can't easily explain why cases of (say) insanity differ from everyday sane criminal wrongdoing. But, more fundamentally, determinism entails that I cannot but write this, you cannot but be (un)persuaded by my arguments, and the criminal justice system cannot but do what it will do. If nothing can change the course of fate, then it will do no good (or ill) to worry when we should (not) hold others culpable.

> There's no free will, says the philosopher
> To hang is most unjust
> There's no free will, assent the officers
> We hang because we must.

Ambrose Bierce, *The Collected Works of Ambrose Bierce: 1909-1912* (Cornell 2009) via Scott Aaronson, 'The Ghost in the Quantum Turing Machine' in S Barry Cooper and Andrew Hodges (eds), *The Once and Future Turing* (Cambridge 2016) 203.

[97] Both Hard Determinism and Libertarianism are minority positions: David Bourget and David Chalmers, 'What Do Philosophers Believe?' (2014) 170 Philos Stud 465, 476: 'Free will: compatibilism 59.1%, libertarianism 13.7%, no free will [hard determinism] 12.2%, other 14.9%.'. I don't find libertarianism in this form very plausible for much the same reason as I doubted OIC*: both reach empirical conclusions via a priori/conceptual premises.

Compatibilism of this sort apparently contradicts Conformability. But that is the case only if both (10) and (11) refer to Φing with the same content. This is not a given. A compatibilist might claim that there is some *conceptually possible* Φing such that we would be culpable for Φing when we cannot but Φ. Nonetheless, they can add, that conceptual possibility isn't realised. For all *actual* Φing, it remains the case, per Conformability, that we are not culpable for Φing if we cannot but Φ. Or the apparent contradiction can be avoided from the other direction by slightly weakening Conformability to say only that *for most Φing*, we are not culpable for Φing if we cannot but Φ. We may yet be culpable where we cannot but Φ in exceptional cases. (Albeit, not so many cases as to undermine the spirit of Conformability.) Once again: Conformability and Compatibilism conflict only if they refer to identical Φing. Still: most compatibilists would find little comfort in these manoeuvres. The spirit of the compatibilist argument, not simply premise (10), is to insulate ordinary free will and culpability attributions from deterministic scepticism about the possibility of doing otherwise. It would fail this task if compatibilism was limited to some subset of merely conceptually possible Φing. Indeed, if they are to accept determinism arguendo, then compatibilists should endorse a stronger premise, according to which for *all* Φing we can be culpable for Φing even if we cannot but Φ. (Insofar as that Φing is a possibly culpable action.)[98] It would be a thin compatibilism indeed to cede all our practical culpability judgments to Conformability.

I noted at the outset that my renditions of these arguments would be but sketches. But this is especially true of Compatibilism. Of its two premises, Libertarians reject Determinism ((6)) while Hard Determinists reject Free Will ((9)). Meanwhile, PAP has independent intuitive support. This, perhaps, is why ordinary people (apparently) intuitively accept incompatibilism.[99] Many compatibilists, therefore, offer an independent argument to reject PAP. They typically rely on the following kind of case:

> *Locked*: D is watching his child in the pool through a window. The child starts to drown, which D realises. But D makes no attempt to save the child, who dies.

[98] As we saw above (fn 85), people disagree about what this caveat amounts to: Darwall says that only wrongful Φing can be culpable, while others reject this restriction.

[99] Shaun Nichols, *Bound: Essays on Free Will and Responsibility* (Oxford 2015) ch 2 for a survey and discussion.

> Incidentally, the doors were locked from the outside such that D could not possibly have saved the child.[100]

Many intuit both that (a) D is culpable for letting the child drown, and (b) D could not have done otherwise than have let the child drown. Thus, it seems to follow, D is culpable for not Φing notwithstanding that D could not Φ. Thus: PAP/Conformability are false. But the problem with this case is again the slippery content of Φ. For plausibly D was culpable for failing *to attempt* to save the child, or for failing to put in place precautions, but not culpable for failing *to save* the child. And, if that intuition is plausible (as I find it), then the obvious explanation for why D is culpable for failing to try, but not for failing to succeed, is precisely that D could have tried, but could not have succeeded. Thus: PAP/Conformability is not falsified.

To avoid my retort, the most well-known counterexample to PAP invokes a stronger counterfactual:

> *Frankfurter*: D of his own volition steals a sausage. But had D decided not to steal it, with certainty Frank would have manipulated D such that D would have chosen to steal it.[101]

The difficulty with this response lies in how we flesh out the relevant counterfactual manipulation. Frank can easily manipulate D by threatening to kill him. But then it's plausible that D can no longer do other than steal the sausage.[102] At any rate, the duress means that D is no longer culpable, and so we don't have a case in which D could do otherwise and yet is culpable. If Frank instead hypnotises D, then plausibly D is no longer culpable in virtue of not performing *any* action: instead, he was a mere automaton. Many who use the example, therefore, rely on a much stronger form of manipulation: something like a precise neurological manipulation to ensure that D 'chooses' to steal the sausage through the same neural circuitry that brings about any other action.[103] But I doubt that this form of manipulation fares any

---

[100] The example is adapted from Tadros, *Criminal Responsibility* (n 2)) 62ff, itself a variant of an example from John Locke, *An Essay Concerning Human Understanding* (1690, Peter Nidditch ed, Oxford 1975) ch 21 para 10.

[101] Based on Frankfurt, 'Alternate Possibilities' (n 93).

[102] This can be doubted: whether (some tokens of) duress literally deprives the duressed party of the ability to do otherwise is controversial. As I discuss in the next chapter, I think a wide reading of what it means to be incapable is plausible.

[103] Frankfurt suggests all three of these forms of manipulation: Frankfurt, 'Alternate Possibilities' (n 93) 835-6. The most famous sci-fi neurological manipulation cases are from Derk Pereboom, in his

better than hypnosis. For this neurological intervention simply seems like hypnosis with bells and whistles. D's choice may *resemble* a 'free' choice, but that appearance is simply the result of the complex aetiological masquerade. He is simply a very precisely manipulated automaton. And automatons are not culpable.

For this reason, I doubt that the standard criticisms of PAP succeed.[104] Naturally, this isn't the final word. Philosophers are broadly split as to whether some version of PAP is workable.[105] But rejecting PAP/Conformability doesn't undermine all of my claims in this chapter. For even those philosophers who reject PAP still accept that some incapacities preclude culpability. Harry Frankfurt, for example, accepts that one is not culpable if one Φed *only because* one could not do otherwise.[106] In effect, then, he simply adds a limitation to PAP: one is not excused for doing what one could not avoid if one was in some way favourably disposed towards one's conduct. But this is probably too strong. Imagine that Frank's manipulation is only strong enough to ensure that D steals something that D independently loves. D's love of sausages is then a necessary condition for the success of Frank's manipulation. It then seems plausible to say that D 'stole' the sausages because (1) of Frank's manipulation, *and* (2) because of D's love of sausages. Thus: it's false that D stole the sausage *only* because he could not do otherwise. But this is the wrong result. D's love of sausages in no way renders him more culpable than D2, identical except for disliking the sausages, who escaped the manipulation (and the crime) only because of that difference.

There are other ways to interpret Frankfurt's tweaked version of PAP, and the literature contains many such versions. My point is simply that those who profess to reject PAP usually add only minor caveats, and uniformly endorse a functionally similar principle according to which some incapacities exculpate by interfering with the sufferers' capacity to avoid violating norms. These accounts differ in their positive prescriptions for what it would take to have the relevant 'regulative control' to render one responsible. But they converge on a relatively settled

---

'Determinism *al Dente*' (1995) 29 Nous 21; *Living Without Free Will* (Cambridge 2001); and *Free Will, Agency, and Meaning in Life* (Oxford 2014) ch 4. Such cases are discussed in Victor Tadros, *Wrongs and Crimes* (Oxford 2017) ch 5.

[104] There are other critiques: Maria Alvarez, 'Actions, thought-experiments and the Principle of alternate possibilities' (2009) 87 Aust J Phil 61; Ralph Wedgwood, 'Rational 'Ought' Implies 'Can'' (2013) 23 Philosophical Issues 70, 74-76.

[105] For a roll call for and against, see Michael McKenna and Justin Coates, 'Compatibilism' in Edward Zalta (ed), *The Stanford Encyclopedia of Philosophy* (Winter 2018 Edition) fn 18, and for discussion see David Widerker and Michael McKenna (eds), *Moral Responsibility and Alternative Possibilities* (Ashgate 2003).

[106] Frankfurt, 'Alternate Possibilities' (n 93) 838-839.

picture as to which incapacities *vitiate* the necessary control.[107] And that settled picture typically involves something like Conformability.

That concludes our whistle-stop tour of the free will literature. In brief: both Hard Determinists and Libertarians accept Conformability but differ as to its implications. Some Compatibilists split on the question (under the guise of PAP), but those who reject it tend to endorse a functionally similar principle. The threat to Conformability from the free will literature is far from insurmountable.

One final point. I doubted both of the major incompatibilist theses and with it the most natural philosophical allies of Conformability. You might wonder what kind of view permits Conformability without sliding into either Hard Determinism or Libertarianism. Here, then, is a sketch of such a view.

Christian List has recently argued for a non-metaphysically extravagant explanation of how it is that we can have the power to do otherwise than we actually do in a deterministic world. His strategy is similar to the one I noted when introducing compatibilism. There I pointed out that the relevant 'Φing' might differ between different premises in the compatibilist argument. List claims that we have the ability to do otherwise if multiple actions are open *for an agent*; an *agential ability*. By contrast, determinism is a thesis about the *physical* state of the world (or, more accurately, how the microphysical state of the world is affected by dynamical laws). Crucially, List claims that *physical* determinism is compatible with an *agential* ability to do otherwise. How is that possible? Unlike traditional Libertarians, he needn't invoke some special power of agent-causation, nor deny that agential abilities supervene on physical facts. Rather, he points out that the psychological states that determine one's agential abilities are *multiply realizable*. That is: different microphysical facts ('microfacts') could determine identical mental states. It follows that two versions of D might be *psychologically* identical at *t1* yet differ in the action they perform at *t2* as a result of their divergent microphysical states dynamically realizing different psychological states. Thus: physical determinism does not imperil a psychological ability to do otherwise, and that is the ability to do otherwise that matters.[108]

---

[107] See fn 81.

[108] Christian List, *Why Free Will Is Real* (Harvard 2019) ch 4. Multiple realizability of psychological states is a complex topic in the philosophy of mind; I leave aside those worries. See John Bickle, 'Multiple Realizability' in Edward Zalta (ed), *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition).

Unfortunately, I don't think List's argument quite succeeds. He emphasises the fact that 'different domains of enquiry describe the world differently' and that 'logically speaking, physical and agential determinism are independent'.[109] But I'm not convinced that his agential ability to do otherwise is sufficient to avoid the usual incompatibilist worries. For consider:

> *Predictor*: Predictor is an advanced intelligence which can analyse the microphysical state of the world to understand the supervening psychological states of human agents. Thus: Predictor can reliably predict which action an agent will perform despite that action being psychologically indeterministic.

Imagine that Predictor predicts your entire future to the minutest of detail by analysing your microphysical state. In this fatalistic world, List's ability to do otherwise is of cold comfort. We look free only so long as we don't look too closely.

But there is a way to salvage the view, based on an argument from the quantum computing theorist Scott Aaronson.[110] The science-fictional flavour of my *Predictor* counterexample is common to many incompatibilist arguments, as we saw with the futuristic neuroscientists capable of causing others to have precisely calibrated psychological states which lead them to perform the neuroscientists' desired actions.[111] Both Predictor and the futuristic neuroscientists invite an image of something like today's algorithms and scientists. But today's algorithms and scientists are no perfect predictors.[112] Some chalk up their predictive failures to free will.[113] But such attributions of free will are vulnerable to future predictive improvements.[114] Predictor and the fantastical neuroscientists are therefore used as plausible

---

[109] List, *Why Free Will Is Real* (n 108) 89 and 97.

[110] Aaronson, 'The Ghost in the Quantum Turing Machine' (n 96)

[111] Cf Alvarez, 'Actions, thought-experiments and the Principle of alternate possibilities' (n 104) on the propriety of attributing *action* to such agents.

[112] Obviously we can predict *choice-based* scenarios very well if there's a strongly preferable choice available. But that doesn't undermine free will. That would require *non-choice based* predictors, such as genetics. Current genetic studies, for instance, can often explain up to 0.4 of the variance in certain traits.

[113] Eg Bryan Caplan, *Selfish Reasons to Have More Kids* (Basic 2011) 82-83.

[114] Compare astronomy. Modern astronomers can predict more than Kepler who could predict more than Copernicus, who could predict more than Mayan astronomers. Copernican astronomy was imperfect. They could have attributed unexplained phenomena to unexplainable divine will. But that would have been premature: modern astronomy has more or less filled the explanatory gap. Could not future scientific developments permit far more accurate prediction, and thus leave us vulnerable to the illusion of free will being shattered?

possible future improvements that would reveal free will to be illusory.[115] Now, I have no general objection to fantastical examples. But they can mislead.[116] For the world, and agents in it, are not necessarily amenable to this kind of prediction. There might remain, in the terminology of economists, a residual *Knightian* uncertainty about which actions one will perform. That is, uncertainty even about the probability distribution that (say) D will Φ. Some Knightian uncertainty might too be resolved by further progress. But not if there is some source of *in-principle* physical Knightian uncertainty. According to Aaronson, this happens to be the case.[117] It follows that our actions necessarily cannot perfectly be predicted. If we cannot know the precise causal mechanism driving our actions, then it cannot be known which choice we will take. Predictor cannot exist. This leaves open a kind of *epistemic* libertarianism:

5) We are not culpable for Φing if we cannot but Φ (*Conformability*)
9*) It is not provable that not: [sometimes we can but Φ] (*Epistemic non-determinism*)
8*) It is not provable that we are never culpable for Φing (*Epistemic free will*)

This claim is weaker than standard Libertarianism. It doesn't claim positively to demonstrate that we have free will. It simply claims that free will cannot be disproven (in a certain way). But it is stronger than List's compatibilism. For it doesn't rely on an ability to do otherwise that vanishes upon scrutiny of its microphysical determinants. Rather, it denies the possibility of scrutiny of those microphysical determinants.

Admittedly, this is not a whole-hearted defence of the freedom to do otherwise. And it relies on the truth of technical details of which I'm not qualified to judge. But who expected the problem of free will to be easy? For now, this seems to me as reasonable a view as any, and it is a view that doesn't undermine Conformability.

---

[115] Some philosophers are content to call themselves compatibilists so long as this illusion (if it be an illusion) gets good results (eg Daniel Dennett, *Elbow Room: The Varieties of Free Will Worth Wanting* (Oxford 1984)). I am not one of them.

[116] A notorious example is John Searle's Chinese Room. (If a man in a room uses a codebook to connect mandarin inputs with mandarin outputs, and thereby (unwittingly) produces a conversation in Chinese, does he thereby *understand Chinese*? (Implication 1: No. Implication 2: Artificial intelligence cannot be the real deal).) But, as others have pointed out, the desired intuition is undermined once we appreciate that the relevant input-output codebook might be planet-sized and the man performing trillions of operations per second. Some form of emergent intelligence no longer seems so outlandish. Daniel Dennett, *Intuition Pumps and Other Tools for Thinking* (WW Norton 2013) ch 60.

[117] Aaronson, 'The Ghost in the Quantum Turing Machine' (n 96).

Conclusion

If I'm right, incapacities exculpate if and insofar as they entail a master incapacity to conform to the requirements of the relevant norm. But how prevalent is this incapacity? I turn to that question in the final chapter.

6 How Incapacities Exculpate

In the last chapter, I defended Conformability: the claim that incapacities exculpate if they entail an incapacity to conform to the relevant norm. But this might be either too harsh or too lenient. It might be too harsh as there are some incapacities which seem exculpatory even though they do not result in the sufferer being literally *incapable* of conforming to a norm. It might be too lenient as incapacities are sometimes hard to distinguish from character traits. If some alleged incapacities are merely character flaws, then to exculpate on that basis is too lenient (§1). These normative critiques of Conformability rely on particular accounts of what it means to be incapable. The first understands a capacity to Φ to mean that Φing is possible for some agent. The second understands some capacities as character traits. Which is correct?

That depends on the meaning and metaphysics of capacity ascriptions. Capacities are standardly analysed via a *counterfactual analysis*. I discuss various problems with such analyses and potential solutions to those problems. I conclude that the problems carry the day. Capacity ascriptions which rely only on metaphysical claims, on the counterfactual analysis, are largely indeterminate. They might be vague and ambiguous or else precise but useless for attributing culpability (§2). If we want to use incapacity ascriptions to attribute culpability, then, we will have to provide the specifications for such claims ourselves. I finish by suggesting that the law could fill in these metaphysical gaps (§3).

1 Two normative critiques

The first critique says that Conformability is too harsh. Some intuitively exculpatory incapacities don't seem to render suffers literally incapable of conforming to a norm. Conformability would then deny that these incapacities exculpate. That seems like the wrong answer. It seems too harsh.

Consider infancy. A seven-year-old child, D, microwaves her hamster, worried that it was feeling cold, or simply being curious as to what would happen. No criminal liability arises for

D in virtue of the infancy defence. This is morally appropriate, we might think, in virtue of the fact that she didn't understand the nature of her actions or didn't understand that it was wrong. But perhaps D had it within her to reflect upon the likely consequences of putting a hamster in the microwave. Or perhaps with some gentle clarification (or stern warnings) she might have come to appreciate the moral problem with microwaving hamsters. Perhaps she could have come to those conclusions on her own. There was nothing preventing her from so concluding. It seems perfectly natural, then, to say that she *could have* conformed to the (moral and legal) norm not to harm her hamster in the way she did. Conformability would then deny that she is exculpated in virtue of epistemic or normative incapacity.[1] But this, intuitively, is the wrong answer. D was too young fully to appreciate the (moral) nature of her conduct. She is morally exculpated on the basis of that incapacity. Schematically,

Too Harsh

1) D had the capacity to conform to the relevant norm
2) D was exculpated on the basis of incapacity
3) ∴ Some incapacities are exculpatory even though the actor did not lack the capacity to conform to the relevant norm

This conclusion denies what Conformability asserts: that an incapacity is exculpatory (if and) only if the actor lacked the capacity to conform to the relevant norm. If *Too Harsh* is right, Conformability is underinclusive as to the set of exculpatory incapacities.

Let's work through the argument backwards. The conclusion, (3), looks valid. But it isn't very strong. It says only that 'some' incapacities don't fit Conformability's story. But Conformability might yet explain the vast majority of exculpatory incapacities. It might only need a minor tweak to avoid such counterexamples.[2] But the critic shouldn't accept this response. For almost all incapacity rules, it seems unlikely that they make conformity literally impossible. As with the infancy example, almost all insane people, those under duress, those harbouring unreasonable beliefs—almost all *could* avoid violating the law. Put a gun to their head and many will conform.[3] If the argument is correct, then Conformability would struggle to explain the vast majority of the incapacity rules.

---

[1] This doesn't preclude a non-incapacity-based excuse, albeit none seem to apply.

[2] See my response to those who tweak the principle of alternate possibilities in ch 5 §6.

[3] Virtually no incapacities leave sufferers literally incentive-blind. Marin Sardy describes how her severely delusional schizophrenic brother Tom ended up in jail, a 'hard, harrowing place for him'. Yet

Alternatively, then, we might deny premise (2). Perhaps D was not in fact exculpated on the basis of incapacity, or anything else. Seven-year-olds generally understand (roughly) how microwaves work. While some rough or uncoordinated play is perhaps to be expected, microwaving pets is not. Children of her age can be expected not to do so. They are culpable for doing so. The doli incapax defence was abolished in part for this reason: because (slightly older) children were thought to be culpable for their failings.[4] As D had the capacity to conform to the relevant norm, Conformability says that she is not exculpated on the basis of incapacity. But she isn't exculpated on the basis of incapacity. Hence, she presents no counterexample to Conformability. (For reasons discussed in chapter 5, a defence might yet be apt for the *class* of 7-year-olds even if its rationale doesn't apply to D). The trouble with this response, however, is that it seems too harsh. The intuitive judgment that D is culpable relies on the fact that children *of her age* can conform to the norm not to microwave hamsters. That they can be expected not to do so. But this is a fact about D's peers. Not about her. If we learned that she was not only young but also developmentally delayed, we might feel less inclined to intuit her culpability. But the pace of child development is sufficiently heterogeneous that she might very well fall into the subset who genuinely lack certain capacities relevant to her culpability despite the fact that her peer group do not. Given a fuller specification of her capacities, it seems very plausible that she might not be culpable, despite first appearances. In that case, we ought not to deny premise (2).

In other words, there is some sleight of hand involved in reaching that culpability judgment. The argument moved from a claim about the capacities of the child's peer group, of what can be expected of them, to a claim about her in particular. It moves from the plausible view that they could conform to the anti-hamster-microwaving norm to the view that she could also conform. But that inference has only weak inductive support. Perhaps she could not conform. In effect, this is to deny premise (1). However, when presenting the case, I noted that D probably had it within her, given some reflection and nudging, to infer the outcomes of her actions and their moral valence. If those steps were taken—steps that were within her control—then she could easily have conformed to the relevant norm. Indeed, even if those steps were not taken, there was nothing physically barring her from not placing the hamster in the microwave and turning it on. Doing nothing was clearly within her power. It seems

---

he managed to avoid going back for a long time, despite not knowing why he was imprisoned, and feeling impelled to perform actions which would amount to offences (like 'rescuing' people). *The Edge of Every Day* (Pantheon 2019) 173, 231.

[4] *C (a Minor) v DPP* [1996] AC 1 (Lord Lowry). Cf the discussion of normality in the last chapter.

that we cannot deny (1). This response, however, contains a whole host of assumptions about what it means to be incapable. Indeed, those assumptions have buttressed all of the responses I considered above. Each claim and each rebuttal relied on a particular, implicit conception of what it means to be incapable of some action or outcome. I noted that *if certain other counterfactuals were present* (like reflecting or being nudged), or *if the child was like her peer group*, then she could have conformed to the norm not to microwave hamsters. I noted that *if a gun was placed to their head*, most people are capable of refraining from certain conduct. But why are these the relevant counterfactuals used to determine when an agent is capable of (refraining from) some action? Why not take the agent as they are, not how they might have been? An answer: this would implausibly restrict the scope of our capacity attributions. If I fail to notice a risk, or fail to appreciate why a noticed risk matters, then in some sense I am not capable of responding properly to that risk. For to respond properly would entail adverting to the risk or its importance. But these failures are the very definitions of negligence and recklessness, respectively. Surely we ought not to draw the boundaries of our capacities so tightly as to define out of existence these standard modes of culpability? But this is to get ahead of ourselves. My point, for now, is neither to endorse nor to refute *Too Harsh*. It is instead to point out its dependence on these difficult and controversial questions as to the very nature of what it means to be incapable. To resolve whether *Too Harsh* succeeds requires an account of what it means to be incapable.

Let's turn to the second critique, which says that Conformability is too lenient. This critique follows on from the last thought above. I suggested that if we define capacities narrowly then it might be the case that standard modes of culpability would be categorised as incapacities. Then we would have a set of *culpable* incapacities. Some culpable incapacities might include the incapacity to conform to a relevant norm. Conformability would wrongly classify these culpable incapacities as exculpatory. Thus: Conformability is too lenient.

Consider a lazy friend or associate. You might have considered whether they were in fact culpably lazy, or else, on more charitable days, whether they were not culpable in virtue of being genuinely incapable of producing the requisite effort. The second critique claims that, in some contexts, this is a false dichotomy. Being lazy simply *is* being incapable of producing the requisite effort. That is just what it *means* to be lazy. It would be inappropriate to exculpate on the basis of this culpable character flaw. It would be far too lenient. Simplified:

Too Lenient

4) Character flaws are culpable
5) Character flaws are incapacities
6) ∴ Some incapacities are culpable
7) If Conformability is correct, incapacities are exculpatory if (and only if) the actor could not conform to the relevant norm
8) Sometimes, character-flaw-incapacities both (a) are culpable and (b) include or entail the incapacity to conform to the relevant norm
9) ∴ Conformability is not correct

The sub-argument in (4)-(6) is reasonably plausible. We equivocate as to whether our friend is culpably lazy or exculpated on the basis of incapacity. If these descriptions refer to one and the same phenomenon, as premise (5) asserts, then we might jump off the fence and consider our friend culpable for their laziness-incapacity. Unfortunately, an opposing argument is also reasonably plausible:

5) Character flaws are incapacities[5]
6*) Characterological features attributable to incapacities are not culpable
10) Character flaws are characterological features
4*) ∴ Character flaws are not culpable

Rather than concluding that our friend is culpably lazy, we might alternatively consider them exculpated in virtue of their incapacity. Intuitive plausibility will only get us so far when evaluating *Too Lenient*. We need to consider whether the premises themselves are well supported. We need to understand the relationship between character flaws, capacity, and culpability.

The key premise is (5), according to which character flaws are incapacities. It is suggested by John Gardner, who claims that

> the extent of someone's capacity for courageous action at time *(t)* is no more and no less than the extent of her courage at time *(t)*. Apart from that courage, there is no further something about her, at time *(t)*, which can intelligibly be described as a capacity for acting courageously at that time.[6]

[5] We could weaken this premise (and, mutatis mutandis, the subsequent argument) to '*Some* character flaws…'. This clarification is noted in premise (8).

[6] John Gardner, 'The Gist of Excuses' in *Offences and Defences* (Oxford 2007) 127.

But this is not obvious. Compare two students. Both achieve a C grade. The first underachieved. She was a bright student who failed to live up to her capacities. The second overachieved. He was a struggling student who pulled out all the stops. It seems perfectly coherent to say that the first demonstrated (prima facie) educational vice, while the second demonstrated (prima facie) educational virtue.[7] This is coherent because their differing capacities part-constitute what it means for their actions to be virtuous.[8] It is courageous for the shy student to speak in front of the class precisely because they lack the capacity to do so with ease. If our capacities part-constitute the *standards* by which to judge character, then our characterological features cannot *themselves* be capacities. Otherwise, the standards would collapse into their exercise. Thus (5) seems mistaken.[9]

Gardner defends (5) by claiming 'that some capacities—virtues of character—are capacities that one does not possess unless one possesses the matching propensities….[Virtues-of-character-]capacities…exist only in their exercise'.[10] The idea is that our character is composed of a cumulative tally of character-constituting actions. There is nothing to being courageous, the thought goes, aside from performing, or being disposed to perform, courageous actions.[11] Otherwise, we must conclude—falsely, thinks Gardner—that some are courageous despite never having acted courageously. But note the equivocation between *performing* and *being disposed to perform* some action.[12] Imagine that our two students have to

[7] Only prima facie, as she may have been justified, excused, etc for her 'failing'.

[8] Barry Mitchell says that virtue must be a 'manifestation of one's personality'. (Discussed in Gardner, 'Reply to Critics' in *Offences and Defences* (n 6) 265.) Gardner replies that our personality is simply our *tendencies of conduct*, which are simply our *capacities*. But it is not clear that Gardner actually subscribes to this view. For example, he writes that 'people [cannot] ever be thought to have a *capacity* for courage when, in their actions, they manifest nothing at all of the courageous mentality' (127). It is phrased as a rhetorical question, but the point is as quoted. (Emphasis added to 'manifest'). It seems that there is *something* to be manifested beyond our actual actions, viz our 'courageous mentality'. Capacities seem to be defined as 'tendencies *plus* mentality'.

[9] Dictionaries define courage both as seeing no fear and as overcoming it. Compare dictionary.com: 'the quality of mind or spirit that enables a person to face difficulty, danger, pain, etc., *without fear*…' to the Oxford English Dictionary: 'The ability to do something *that frightens one*.' (Emphases added). Nelson Mandela accepted the OED version: 'I learned that courage was not the absence of fear, but the triumph over it.' Nelson Mandela, *Long Walk to Freedom* (LBC 1994) 542. Despite its Aristotelian flavour, Gardner's view finds only lukewarm support in Aristotle. While Aristotle thought those with no fear were *most* virtuous, he still considered the merely continent *somewhat* virtuous. Philippa Foot, *Virtues and Vices* (California 1978) 8-12; Victor Tadros, *Criminal Responsibility* (Oxford 2005) 314-315.

[10] Gardner, 'Reply to Critics' (n 6) 263. See too 'The Gist of Excuses' (n 6) 126.

[11] Gardner, 'The Gist of Excuses' (n 6) 126-127. What counts as courageous is not the mere mechanism of the action, but also depends on the spirit in which it is done, etc.

[12] Gardner sometimes notes the equivocation. He writes 'A soldier should be more level-headed than most (and therefore, if that really is something different, needs a greater capacity for level-headedness

perform a song they have never sung before. Imagine, indeed, that they have never sung anything before. As it turns out, the first sings beautifully, the other not so much. But was the first a good singer before they had sung? This seems plausible. Innate talent varies. The first was the better singer before either opened their mouths. And such a judgment can be made without the benefit of hindsight based on hypotheticals, counterfactuals, or other forms of evidence. (Perhaps one student's parents are professional musicians). Now, virtue is composed both of talent and graft. By analogy, then, it seems plausible that one student might have been lazier or more industrious than the other even before they had a chance to *evidence* that trait. But that evidence need not consist of exercises of the very trait in question. Once again, it might be counterfactual. The need for exercise seems epistemic rather than constitutive. Thus, it seems to me that, *pace* Gardner, virtue of character capacities do not exist only in their exercise. Capacities, including Gardner's character-capacities, can be unexercised. Character judgements would be justified by the fact that the one being judged was *disposed* to act in accordance with the character trait. Now, dispositions are a fraught philosophical topic. But talk of propensities, tendencies, and dispositions usually presupposes that they can exist without exercise.[13] Then it remains perfectly reasonable to say that virtues and vices relate to our actual actions whereas our capacities—something like the set of possible unactualised actions—are an entirely different matter.[14] Thus, again, it looks like (5) is false. And if (5) is false, *Too Lenient* doesn't work.

---

than most)…' and also '[I]f she is incapable of such loyalty (or, in my plainer terms, if she is not loyal enough).' ibid 127.

[13] Some philosophers analyse statements of probability as statements about propensities. These propensities are taken to be some kind of metaphysical feature of an objectively homogeneous set of *hypothetical* outcomes. Karl Popper, 'The Propensity Interpretation of Probability' (1959) 10 BJPhilSci 25; Antony Eagle (ed), *Philosophy of Probability: Contemporary Readings* (Routledge 2010) ch 26. I am not saying that propensity analyses of probability are correct; indeed, I doubt it: see Anthony Eagle, 'Twenty-One Arguments Against Propensity Analyses of Probability' (2004) 60 Erkenntnis 371. The point is simply that 'propensities' (etc) don't involve manifestations of action. As Charles Pierce puts it, a 'statement… [about the] probability [of a die throw]…[means] that the die has a certain "would-be"; and to say that the die has a "would-be" is to say that it has a property, quite analogous to any *habit* that a man might have.' (Quoted by Alan Hájek, 'Interpretations of Probability' in Edward Zalta (ed), *The Stanford Encyclopedia of Philosophy* (Winter 2012 Edition). See too Popper, 'The Propensity Interpretation' at 30.)

[14] Strictly speaking, Gardner only endorses this actualist reading of character-capacities. But he leans towards actualism about capacities in general in 'Reply to Critics' (n 6) 259 fn 23, referring to John Gardner, 'The Mysterious Case of the Reasonable Person' (2001) 51 UTLJ 273, 281ff.

There are other possible replies and objections to Gardner's view of the relation between excuses and incapacities.[15] Once again, my point isn't categorically to refute them. It is rather to note that *Too Lenient*, an apparently normative critique of Conformability, relies for its plausibility on controversial premises as to the nature of capacities. The real action in both normative critiques is this metaphysical question.

2 Counterfactual analysis

The aim. The incapacity rules are supposed to cleave

> *Culpable*: D ran over a child with a truck. He could have stopped.

from

> *Excuse*: D ran over a child with a truck. He could not have stopped.

To do this they need to make sense of what 'could (not)' means.[16]

Capacities refer to some verb: a capacity *to stop the truck*, a capacity *to distinguish right from wrong*, a capacity *to conform to a norm*. When talking of capacities generically, substitute those verbs

---

[15] Gardner claims that 'For a person who sees the world through genuinely courageous eyes, there is … no room for any motivational obstruction between understanding why a situation calls for courage and acting courageously.' Gardner, 'The Gist of Excuses' (n 6) 127. This implies either (1) a denial of the possibility of akrasia, or (2) that akratic action is necessary culpable. Both views are controversial. This also seems to imply the claim that understanding and acting are processed by the same physical mechanisms. But, empirically, the neurological mechanisms which generate actions (including those deemed virtuous) are (partially) distinct from those which govern motivation. Gardner disavows that claim: 'I said nothing about why people are brave, or what brings it about that they react bravely today and without any bravery tomorrow, or similar questions of aetiology. I have no views on these matters which… require empirical investigation.' Gardner, 'Reply to Critics' (n 6) 264. Instead, Gardner's view may be that 'virtue' is denoted by the combination of the mechanisms of understanding and acting. But this simply stipulates away the difficulty presented by cases in which people (would) have the requisite virtues but for suffering a defect of motivation (e.g. from depression, akrasia, or whatever). Consider Tourette's syndrome. Assume that the Tourette's sufferer has the general capacity to avoid swearing (perhaps the condition only manifests under stress. Few would say that D's tendency is identical to a lack of actual virtue. (See eg Judith Buse et al, 'The Modulating Role of Stress in the Onset and Course of Tourette's Syndrome: A Review' (2014) 38(2) Behaviour Modification 184.)

[16] What someone *did* is analytically primitive: D did Φ if and only if it is the case that D Φed (analogously to Tarski's account of truth, according to which 'T is true' if and only if it is the case that T.) This ducks the central question in the philosophy of action, viz how to distinguish 'D did Φ', a claim about the *action* of Φing, from a claim about the mere *event* of Φ occurring. What is the relationship required between the agent, D, and the event, Φing, such that it was D *who Φed*? Some analyses of action rely on an account of the agent's capacity, eg Φing is an action only if the agent *could* bring about the relevant output. This is another reason to focus on the meaning of capacities. Judges in criminal cases are concerned with evaluation of capacities after the fact: 'D could have' rather than 'D can'. But I assume temporal symmetry, such that 'D could have Φed' is true at t2 if and only if it is true at t1 that 'D can Φ'.

with 'Φ'. The incapacity rules aim to account for those cases where D could not (relevantly) Φ.[17] If Conformability is correct, then the relevant 'Φ' for the incapacity rules is 'conform to the norm'. But an analysis of the meaning of capacities in general can remain aloof of that particular account of the relevant incapacity. Ideally, an account of capacities will draw the boundaries of what it means to be (relevantly) (in)capable to coincide, roughly, with the set of cases in which the incapacity rules actually apply. This criterion of extensional fit will be met only if our account of capacities implies that the incapacity rules deny or defend only a minority of (acts that constitute) offences. But fulfilling this criterion of extensional fit is easier said than done.

The problem. Our trucker is culpable because he ran over a child when he could have stopped. But, as many emphasise in the free will literature, it's not clear when or if he could have done otherwise. Once we scrutinise the actual causal chain leading to the trucker's act or omission, it is less obvious what it means to say he could have stopped. Say he was grossly negligent: he failed to advert to the risk of hitting the child. Still, we might wonder: *could* he have so adverted?[18] Or he was reckless: he realised the risk, but simply didn't concentrate hard enough to avert the danger. *Could* he have so concentrated?[19] Or he intentionally mowed down the child. *Could* he have resisted the impulse?[20] Even leaving aside global worries about determinism, the answer to each question is not obviously 'yes'. When focusing on the actual causal chain which leads to the trucker's actions, it seems like various things must have been different for the outcome to differ. Some differentiators are external to the trucker (the presence of the child in the road). Others are internal (his perceptions, beliefs, motivations, etc). The trucker might have been capable of doing otherwise in those counterfactual worlds where things were different. But in the actual world, at the very moment of action, it looks

[17] What if did could not but did Φ? Is this not contradictory? That depends. I cannot solve a Rubik's cube in 10 seconds. Yet it's conceivable that through dumb luck I do just that. Hence, it appears that I could not yet did solve the Rubik's cube in 10 seconds. This hints at some difficulties explored below.

[18] Some advocate reforming the negligence standard to include this 'could-have-adverted' condition. Eg Simester et al, *Simester and Sullivan's Criminal Law* (6th edn, Hart 2013) 162-166.

[19] Law Commission, *Criminal Liability: Insanity and Automatism* (Discussion Paper, 2013) [1.35]: 'The term "defect of reason" has been interpreted to mean that for the defence of insanity to operate, the accused's powers of reasoning have to be impaired at the time of the commission of the offence. A mere failure to use powers of reasoning is not enough. Momentary failure of concentration, even where caused by mental illness, is not insanity within the M'Naghten Rules.' (Citing Ackner J in *Clarke* [1972] 1 All ER 219, 221.)

[20] 'The line between an irresistible impulse and an impulse not resisted is probably no sharper than between twilight and dusk'. Per the American Philosophical Association, as cited in RD Mackay, *Mental Condition Defences in the Criminal Law* (Oxford 1995) 115.

like he could scarcely have done otherwise. The notion of 'could have but did not' seems puzzling.

Counterfactual analysis. Compare talk of causation. The trucker caused the death of the child. But what does this mean? Roughly: if the trucker had not driven into the child, then the child would not have died. Lawyers are familiar with the required caveats. But causation-talk is necessarily counterfactual-talk, and this embarrasses no lawyer. So too, then, capacity-talk's dependence on counterfactuals is not necessarily problematic. To say that we did not but could have Φed is to say something about a counterfactual world in which we did Φ.[21] As Austin memorably put it, capacity attributions are 'constitutionally iffy': to say 'the trucker could have stopped' suppresses the qualifier 'if…'.[22] Further, to *analyse* a term requires substituting the term of interest with a different term. Analysing modals requires substituting non-modals.[23] Thus, 'D could Φ' is standardly analysed to mean 'D *would* Φ if {…}'. This strategy is called counterfactual (or conditional) analysis ('CA'). The '{…}' specifies the set of counterfactual conditions under which D would Φ for it to be true that D 'can' Φ or 'could have' Φed. The condition standardly suggested is that D *tried to* Φ. The trucker *could* have stopped if he *would* have stopped if he *tried* to stop.[24] (Necessary conditions external to him—eg having functional brakes—are standardly dubbed the *opportunity* to Φ. The CA is true only if we assume the opportunity).[25]

However, this simple form of CA fails. It claims to offer necessary and sufficient conditions for capacity ascriptions. But there are problems with both claims.[26]

---

[21] John Maier, 'Abilities' in Edward Zalta (ed), *The Stanford Encyclopedia of Philosophy* (Fall 2014 edition).

[22] JL Austin, 'Ifs and Cans' in his *Philosophical Papers* (JO Urmson & GJ Warnock eds, Clarendon 1961).

[23] See the paragraph on possible worlds below.

[24] David Hume, *An Enquiry Concerning Human Understanding* (1748, PH Nidditch ed, Oxford 1978) 73, cited in Michael McKenna and Justin D Coates, 'Compatibilism' in Edward Zalta (ed), *The Stanford Encyclopedia of Philosophy* (Winter 2018 Edition) §3. The strategy is often used in modern work too (albeit not usually on work focused on capacities): see, eg Derek Parfit, *On What Matters: Volume One* (Oxford 2011) ch 11 §38.

[25] As John Gardner puts it, capacities must be something '*about me*'. (Gardner, 'The Gist of Excuses' (n 4) 126.) D might be culpable for not stopping despite lacking the opportunity to stop if D was culpable for inducing the absence of opportunity (eg failing to check the brakes), just as D might be culpable for lacking the capacity to stop (eg due to unconsciousness) in virtue of prior fault (eg getting drunk).

[26] My objections are not novel. See eg McKenna and Coates, 'Compatibilism' (n 24) §3.3; Maier, 'Abilities' (n 21) §3.

Sufficiency. CA says that 'D can Φ' if 'D would Φ if {D tried to Φ, …}. But what if D cannot try to Φ? Consider:

Claustrophobia: I cannot try to enter small spaces.

CA says that I can enter small spaces if I would enter them if I tried to do so. Claustrophobia says I cannot try to enter small spaces. Nonetheless, it's true that *if* I tried then I'd succeed. CA concludes that I can enter a small space. But this is false. I cannot enter small spaces.[27] It's false because there is an incapacity embedded in the counterfactual itself. Thus, the truth of CA is not sufficient for capacity attributions.[28] D must also be capable of fulfilling the counterfactuals *within* the analysis. CA can be modified to avoid this objection, by changing

1) D would Φ if D tried to Φ

to

2) D would Φ if (D tried to Φ) and (D can try to Φ)

Incorporating a capacity ascription in the second clause, however, requires further analysis. For we must analyse the 'can' in '(D can try to Φ)'. Using the same strategy, this unpacks to 'D would try to Φ if D tried to try to Φ'. And, of course, we also need to add the rider that 'D *can* try to try to Φ.' All together, then:

3) D would Φ if (D tried to Φ) and ((D would try to Φ if D tried to try to Φ) and (D can try to try to Φ)

This modification has obvious drawbacks. The modification adds an additional can-clause, but that additional can-clause must itself be analysed in terms of further can-clauses. This

[27] Richard Feynman recounts the experience of being hypnotised: 'You're only slightly fogged out, and although you've lost a little bit, you're pretty sure you could open your eyes. But of course, you're not opening your eyes, so in a sense you can't do it... All the time you're saying to yourself, "I could do that, but I won't"—which is just another way of saying that you can't.' By contrast, the experience of sensory deprivation whilst on ketamine created the inverse phenomenology: '[A]lthough I had a feeling of complete disorientation, a feeling of an inability to do practically anything, I never found a specific thing that I couldn't do.' 'Meeeeeeeeee!' and 'Altered States' in *Surely You're Joking, Mr Feynman* (Bantam 1985).

[28] A similar example involving sweets is in Keith Lehrer, 'Cans without Ifs's' (1968) 29 Analysis 29, 32.

results in an infinite recursion of postulated capacities. Some take this to be a decisive objection to CA.[29] But that is a little hasty. Contrast:

> *Weak claustrophobia*: I cannot try to enter small spaces unless I try to try to do it.
> *Strong claustrophobia*: I cannot try to enter small spaces regardless of whether I try to try to do it.

The distinction here is not meaningless. We're all familiar with 'psyching oneself up'. I cannot enter small spaces without further ado: I find it extremely unpleasant. Indeed, the phobia makes it very hard for me even *to try* to do it. But perhaps I can try *to try* by psyching myself up: by imagining that the small space is large, by mentally blocking fears of suffocation, etc. This distinction might allow us meaningfully to distinguish between capacity and incapacity. This could work in two ways. First, we might *stipulate* that counterfactual capacity ascriptions require only two levels of recursion. I can Φ if I would Φ if I tried to Φ and can try to Φ and can try to try to Φ. But, by stipulation, I can *still* Φ if I would Φ if I tried to Φ and can try to Φ and can try to try to Φ *but cannot try to try to try to* Φ. The necessary capacities only run two levels deep. Second, alternatively, perhaps the set of real-world cases diminishes exponentially with the level of recursion.[30] The simple CA correctly distinguishes the majority of cases: I'm not raising my hand right now as I'm not trying to raise my hand, whereas for someone paralysed they would not raise their hand even if they tried to do so. I can raise my hand, they can't. CA gets this right. But it gets a small number of cases wrong, cases like *Claustrophobia*. Adding the second can-clause solves these cases. There still remain some cases where the agent can try to try to Φ but cannot try to try to try to Φ. But the number of such problem cases is minuscule. And they reduce asymptotically to zero as we add levels of corrective can-clauses. Thus: the problem of sufficiency *practically* solves itself with a (practically) non-infinite proliferation of additional can-clauses. But note the limitation to this modification. It remains the case that self-reference is an *analytical* deficiency of any conceptual analysis.

Necessity. CA says that if 'D can Φ' then D *would* Φ when the conditions obtain. But all the conditions to Φ might obtain, rendering D intuitively capable of Φing, despite the fact that D would *not* Φ in some cases. Consider:

---

[29] Alfred Mele, 'Agents' Abilities' (2003) 37 Nous 447, reprinted in Alfred Mele, *Aspects of Agency: Decisions, Abilities, Explanations, and Free Will* (Oxford 2017). (References per the original article).

[30] I mean this loosely: it might not be literally exponential.

Golfer: A good golfer tries to make an easy putt. Unusually, he fails.

Intuitively, all good golfers can make easy putts. CA says that this mean 'A good golfer would make an easy putt if they tried to make an easy putt'. But even good golfers miss some easy putts. In those cases, the counterfactuals of the analysis obtain but D would not Φ. CA then denies that D was capable of making the putt in that instance. But this seems false. The golfer 'can' make the easy putt, failure notwithstanding. Contra the simple form of CA, capacity attributions are robust to some failures.[31] Again, we can try to modify the simple CA to account for such counterexamples. The obvious suggestion is to weaken

1) D would Φ if D tried to Φ

to

4) D *might* Φ if D tried to Φ, or
5) D would *normally* Φ if D tried to Φ

These modifications also have an obvious drawback. They make capacity attributions very vague. If I tell you that I can make the meeting, it's not enough that I *might* make the meeting. We want more certainty than that. Nor is it enough that I *normally* make such meetings. For what is the relevant reference class? What if this case is importantly different? (I *normally* make such meetings, but *this* week I'm in Taipei, or taking a sabbatical, or hospitalised, or…). Judges implicitly resolve these ambiguities whenever they decide that some defendant could or could not Φ. Those resolutions contain an implicit account of the relevant disambiguation. But different (implicit) disambiguations output significantly different judgments across real-world cases. Perhaps judge A uses a 60% probability threshold for incapacity attributions, whereas judge B uses a 90% threshold. Defendants who claim that an incapacity defence applies to them may fall between these bounds. The two judges will therefore decide these cases differently. Judge A will accept that an incapacity defence applies, whereas judge B will not. A defendant's liability turns on the account the judge uses. But without some clearer criteria of correct capacity ascriptions, judges' decisional freedom seems offensively unconstrainted.

Let's take stock. In response to objections to both the sufficiency and necessity of CA, we've proposed two modifications. Instead of the simple CA, we might analyse 'D can Φ' as:

[31] The example is from Austin, *Philosophical Papers* (n 22) 166 fn 1.

6) (D would normally Φ if D tried to Φ) and (D can try to Φ)

Something like (6) *might* work for practical capacity ascriptions in the law. But this is far from an ideal analysis of the concept of capacity. The first clause is very vague, and the second clause contains the very term to be analysed. There have been other attempts to modify CA to defend against these kinds of problems, but none have found much success.[32] These apparent failures of CA, both simple and modified, prompt three responses: actualism, bullet-biting, and disambiguation.

Actualism. The simplest response is to retreat to metaphysical scepticism about unexercised capacities. Regarding *Golfer*, we might accept that it is simply untrue that he has the capacity to sink easy putts which he tries and fails to sink. For if he was capable of doing so, why did he fail? In slogan form, 'D can Φ only if D is Φing or has Φed'. This view, *actualism*, disavows counterfactual analyses of capacities.[33] Regarding our trucker, it says that there is simply no truth in the claim that the trucker could have but did not stop. If he stopped, then ipso facto he was capable of stopping. But if he didn't stop, well, he was not so capable. Scepticism eliminates all our usual talk about unexercised capacities. This is radically revisionary. It could hardly fare worse on the criterion of extensional fit.

Bullet-biting. An alternative to actualism is to accept CA but to reject the force of the counterexamples. We bite the bullet and attribute capacities even in in the face of counterexamples. We say that the Claustrophobic really *can* enter small spaces; that psychological inabilities embedded in the counterfactual conditions are simply to be ignored for the purposes of capacity ascriptions. This response is also highly revisionary. Regarding *Golfer*, CA says that he cannot Φ as he didn't Φ despite fulfilling the conditions of the CA. Much negligent and reckless wrongdoing is performed despite the agent trying to do otherwise. Thus, it appears that much wrongdoing ordinarily presumed to be capable will be reclassified as incapable. Regarding *Claustrophobia*, bullet-biting is less unintuitive. Some deny

[32] Maier, 'Abilities' (n 21) §3.3. This, perhaps, should not be surprising. Gettier's famous objection to the view that Knowledge = Justified True Belief spawned a large literature trying to modify that formula with extra conditions. The attempt was widely considered a failure. One lesson, perhaps, is that tinkering with simple analyses of fundamental concepts might not be a winning strategy. Instead a bigger change of tack might be required.

[33] Sometimes dubbed Megarian actualism after the Greeks who apparently endorsed the view: Aristotle, *Metaphysics Book IX* (Stephen Makin trs, Oxford 2006) ch 3 (aka book IX.3).

that psychological incapacities are incapacities properly so-called.[34] But there are other examples of non-psychological incapacities which apparently contradict CA. Consider:

Blind: D is blind. E throws a frisbee in her direction. D doesn't see it.

Can D dodge the frisbee? Yes, says CA: if D tried to dodge it, she would succeed. But can she try to dodge it? I equivocate. Yes: nothing was stopping her, per se. No: being able to try to Φ requires that Φing is a salient, candidate action. That requires not only certain psychological attributes (an absence of overwhelming phobias) but also simple perceptual attributes (sight). The bullet-biting response here—simply saying that D can dodge the frisbee, blindness notwithstanding—seems, if not wrong, then at least hasty. Bullet-biting has baggage. Notice, however, that these two revisionary bullet-biting moves (in part) cancel out rather than amplify their revisionary implications. Perhaps most negligent and reckless wrongdoing *doesn't* entail the agent trying to do otherwise, but instead is better understood as the agent being *unable to try* to do otherwise. (Whether because they do not advert to the risk, or because they fail appropriately to infer, from adverting to the risk, that they ought to alter their conduct, and (in that sense) 'cannot' avoid acting recklessly). Under this description of their failure, the bullet-biting response simply rejects that their incapacity to try to Φ undermines (ascriptions of) their capacity to Φ. Then the revisionary implications of bullet-biting are constrained. Bullet-biting then fares quite well on the criterion of extensional fit. (Indeed, it does even better if, per the earlier modification, we bite the bullet only for second-order can-clauses (rejecting that an incapacity to try *to try* to Φ undermines capacity ascriptions).)

In the choice between scepticism and bullet-biting, bullet-biting seems clearly preferable. It fits better both with our pre-theoretical intuitions and the law. But it remains imperfect. Perhaps we can tolerate analytical self-reference for practical purposes. But the extreme vagueness of the claim that 'D would *normally* Φ' would make practical (in)capacity ascriptions correspondingly vague. That is a serious problem for the law. This suggests a final response to the problems of CA: disambiguation.

Disambiguation. Recall two analyses of 'D can Φ':

[34] The most well-known proponent is Thomas Szasz, in various books starting with *The Myth of Mental Illness: Foundations of a Theory of Personal Conduct* (Perennial 1961). Many endorse a narrower version of the claim, eg that addictive impulses are not literally irresistible. Bryan Caplan offers this as a distinction between constraints and preferences in 'The Economics of Szasz: Preferences, Constraints and Mental Illness' (2006) 18 Rat. and Soc. 333 (critically discussed by Scott Alexander, 'Contra Caplan on Mental Illness' *Slate Star Codex* 7 October 2015.

1) D would Φ if D tried to Φ

5) D would *normally* Φ if D tried to Φ

CA (1) is vulnerable to counterexamples like *Golfer*. (5) is hopelessly vague. One solution to both problems is to disambiguate between the two when ascribing capacities. Tony Honoré does this in distinguishing what he calls *general* from *particular* capacities.[35] A *particular* capacity relates to a single case. We can (particular) do only what we in fact do. *Golfer* cannot (particular) sink the putt. Likewise, an incompetent golfer can (particular) hit a hole-in-one just in case they get freakishly lucky and do just that. In other words, Honoré embraces an actualist interpretation of (1), but only for one-shot capacity ascriptions. By contrast, a *general* capacity refers to a set of cases. It says that *Golfer* can (general) sink easy putts, as he normally does so; that the incompetent golfer cannot (general) make holes-in-one, as he normally fails. That is, Honoré bites the bullet per (5), but only for capacity ascriptions regarding a set of cases.[36] As we saw above, both (1) and (5) have some intuitive appeal, but apparently fundamental weaknesses. Honoré squares the circle by embracing both—not in contradiction, but in disambiguation. Different analyses suit different purposes. We can use (5) for capacity ascriptions requiring some robustness to failure ('can *Golfer* sink easy putts?'),[37] but use (1) for capacity ascriptions requiring analytical precision ('can *Golfer* sink this very putt?').

Our practical interest in the distinction, however, lies entirely with the concept of general capacities. As I noted at the outset, the job of the incapacity rules is to cleave *Culpable* from *Excuse*. It is to distinguish *between* cases in which an offence was committed from cases in which a (would be) offence was denied or defended on the basis of incapacity. Asking whether a defendant lacked a relevant *particular* capacity is useless: the defendant could (particular) do (all and) only the things they in fact did. So the incapacity rules make sense only on the assumption that they implicate general capacities. Honoré's concept of general capacities has been widely endorsed and emulated by philosophers offering accounts of

[35] Honoré, 'Can and Can't' in his *Responsibility and Fault* (Hart 1999).

[36] ibid 147. Honoré adds a second condition that 'there is nothing in the circumstances to prevent the agent's succeeding on this occasion'. This refers to *opportunities*, per John Gardner, 'Reasons and Abilities: Some Preliminaries' (2013) 58 Am.J.Juris 74.

[37] Daniel Dennett, *Elbow Room* (Oxford 1984) 145: 'Hume speaks of "a certain looseness" we want to exist in our world…the looseness that prevents the possible from shrinking tightly around the actual, the looseness presupposed by our use of the word "can."'

capacities.[38] It has been used (both explicitly and implicitly) by legal philosophers discussing the relevance of (in)capacities to responsibility.[39] But, given that it remains essentially just CA per (5), it appears hopelessly vague.

Precisifying success. One possible solution is to precisify exactly how much success counts as '*normal*' success. Split Φ into success percentages: Φ_1, Φ_2, Φ_3…Φ_{100}. Quantify 'normal' success as (say) 80% success, or Φ_{80}. Thus: a general capacity to Φ = one would Φ 80% of the time (under the relevant counterfactuals) = Φ_{80}. The incapacity rules say that D is exculpated if D cannot Φ in the relevant sense. Applying the newly precisified concept of a general capacity, this means that D is exculpated if he wouldn't Φ_{80} under the relevant counterfactual conditions. The problem, however, is that being able to do something normally, or 80% of the time, is perfectly consistent with being *unable* to do it on a specific occasion. Consider:

> *Golfer 2*: A good golfer tries to make an easy putt. Before hitting the ball, however, he suffers a stroke.

It is true that *Golfer 2* can normally make that putt; can make the putt 80% of the time. But it is false that *Golfer 2* can make the putt on this occasion. Honoré's distinction captures both sides of the story: he can (general) but cannot (particular) make the putt. But which sense is relevant to culpability attributions? *Golfer 2*'s teammates (fans, audience, himself…) might justifiably 'blame' him for missing a putt which he could have sunk. They ought not to blame him for missing a putt which he could not have sunk. Is he to blame or not? It is no answer to say that he could (general) and could not (particular), for we want to know whether he is to be blamed or excused. The correct answer is clearly that *Golfer 2* is excused. But that means the sense of capacity relevant to culpability ascriptions is not general capacities. If not general, then, on Honoré's bipartite division, it must be particular capacities. But particular capacities entail actualism, which entails that no-one can do what they do not do. And that

---

[38] Alan Millar says that capacity ascriptions imply that success is *reliable*, that failures must be '*very rare*': Alan Millar, *Knowing by Perceiving* (Oxford 2019) 128. John Maier says that the agent must be successful in a '*suitable proportion*' of cases. John Maier, 'The Agentive Modalities' (2013) 87 Phil and Phenom Res 1, 15-16. Like Honoré, Maier distinguishes particular from general capacities, albeit the former under the label 'options'. (Though his account of options, unlike Honoré's particular capacities, is itself a modal concept). Fusing these views, Alfred Mele suggests that general capacities require 'sufficiently reliable' success. Mele, 'Agents' Abilities' (n 29) 464. As I'll note shortly, Mele subdivides general capacities. But this minimum success is common to both versions he offers.

[39] Tadros, *Criminal Responsibility* (Oxford 2005) 58, Joseph Raz, 'Being in the World' in *From Normativity to Responsibility* (Oxford 2011) 247.

implies that D would *always* be exculpated for failure on the basis of incapacity. Honoré's distinction might be useful in some contexts, but it cannot furnish us with a sense of capacity fit for culpability attributions.

This problem has been overlooked by some who rely on Honoré's distinction. Victor Tadros claims that

> If the agent is capable of appreciating norms in general [ie has general normative capacity], he is capable of appreciating this norm in the relevant sense [for responsibility attribution].[40]

But *Golfer 2* explains why this is misleading. For Tadros' claim amounts to the following:

> An agent is capable of appreciating norm Φ in the relevant sense for responsibility attribution if they are capable of appreciating norm-type Φ, *albeit not necessarily norm-token Φ*.

The final clause is necessarily entailed by general capacity ascriptions. But making it explicit reveals the problem with relying on such capacity ascriptions when attributing culpability. Being generally capable of appreciating norms is not sufficient to attribute a capacity to appreciate norms in any particular instance for such purposes. Tadros's inference can be drawn only if *this* token norm appreciation is aptly subsumed within the class of general norm-appreciations. It is not apt in cases like *Golfer 2*. Thus, Tadros's claim at best requires supplementation. What makes any such inference apt? That is not obvious.

Honoré himself contributes to this misuse of general capacity ascriptions. First, he claims that particular incapacities (partially) mitigate blame.[41] But all wrongdoing entails a particular incapacity to avoid wrongdoing. If Honoré were right, all culpability for wrongdoing would be mitigated. Honoré (correctly) did not intend that conclusion. Thus, he must have erred in claiming any mitigatory effect for particular incapacities. Second, recall those cases where I can try to Φ but cannot try to *try to* Φ. If I have the general capacity to Φ, Honoré says that this notionally entails that 'an indefinite number of projected "can" [particular] statements have been answered in the affirmative.'[42] But this implies Φ_{100} rather than Φ_{80}. In fact, no such indefinite sequence is necessary when ascribing general capacities. We need only assume

[40] Tadros, *Criminal Responsibility* 137.

[41] Honoré, 'Can and Can't' (n 35)148.

[42] ibid 156.

enough of that sequence is answered in the affirmative to account for Φ_{80} over the total set of cases. Honoré goes on to add that in 'practice, we dispense with this [infinite questioning] and are satisfied if there is no immediate or obvious obstacle to a given choice.'[43] But this is not at all what we are doing. Rather, general capacity ascriptions survive even when some of those projected particular capacity statements fail. The fact that I have the general capacity to Φ does not entail that I can Φ 100% of the time—Φ_{100}—even if trying. Honoré's suggestion, like Tadros's, suppresses those concomitants of general capacity ascriptions which undermine its plausibility for use in culpability ascriptions. Despite Tadros and Honoré's suggestions, then, culpability ascriptions predicated on capacity ascriptions cannot rely only on the concept of general capacities.

One response to the problem of vagueness and the inadequacy of precisifying the success threshold is to offer a more nuanced account of 'normal' success. John Maier, who talks of a 'suitable proportion' of success, claims that this standard is flexible.[44] If we predicate 'Can raise our arm' over the entire day, virtually no-one has the general capacity to raise their arms. Why? Well, given that most of us sleep for a third of the day, we can only raise our arms about 66% of the time, ie Φ_{66}. That falls short of our stipulated minimum success threshold of Φ_{80}. We can solve that problem by flexibly lowering the minimum success threshold to account for such contextual factors. Perhaps, then, we ought to ascribe a general capacity to raise our arms at Φ_{60} rather than $\Phi_{80.}$ (Conversely, perhaps the capacity *to breathe* is ascribed only at Φ_{99+}.) Maier suggests that this flexibility also makes sense of *gradations* of capacity: you and I are both capable of raising our arms, but perhaps you, sleeping less, are *more* capable.[45] Such flexibility need not be restricted to raw numbers. We noted in the last chapter that judgements of normality are sensitive to both descriptive and prescriptive norms.[46] Perhaps 'normal' success in raising one's hand is a function not only of a minimum success threshold but also the quality or appropriateness of the action in question. Imagine a schoolchild who is generally perfectly able to raise their hand, but, due to crippling anxiety, is psychologically unable to raise their hand when asked to speak in class. Given that they're rarely asked in class, we might say naively that they can (general) raise their hand (say, Φ_{85}). But, given the contextual importance of hand-raising in response to

[43] ibid 158.

[44] Maier, 'Agentive Modalities' (n 38) 16-17.

[45] ibid 16.

[46] Adam Bear and Joshua Knobe, 'Normality: Part Descriptive, Part Prescriptive' (2017) 167 Cognition 25-37. I discussed this claim in ch 6 §2 ('Normality').

questions, we might downgrade that capacity ascription to account for the normative significance of the classroom context.

Unfortunately, this response doesn't solve the problem. It might yield a richer account of general capacities by accounting more flexibly for (what counts as meeting) the necessary success threshold. But it does this only by exacerbating the main problem with the use of general capacities: it makes capacity ascriptions extremely vague. For not only is it unclear whether D would normally Φ, it's also unclear how much success even counts as normal, or even—using the hybrid descriptive/normative concept of normality—what counts as success.

Precisifying the counterfactual. Precisifying the necessary success threshold alone won't yield an adequate account of capacities for use in culpability ascriptions. But there are other ways to make the concept more precise. Chief among these is precisification of the relevant counterfactual. Regarding the student who can only sometimes raise their hand, we might distinguish classroom-hand-raising from non-classroom-hand-raising. For plausibly they can raise their hands outside of the classroom but cannot raise their hands within it. Contextual precision allows us to offer more fine-grained capacity ascriptions. Precisifying context is not sharply distinguishable from precisifying success, for the success threshold always presupposes a certain counterfactual context. Some formulations of general capacities therefore straddle the two ideas. Jack Spencer talks of Φing in 'representative attempts', where representativeness is a function both of similarity and success.[47] Alfred Mele's concept of 'promise-level ability' requires not only a high probability of success but also requires that failures are attributable to external factors—a limit on the relevant counterfactual context.[48] By contrast, John Maier keeps the two ideas distinct, in requiring success in a suitable proportion of 'similar cases'.[49] But whether success and context are kept distinct or not, both are precisified to restrict the set of relevant counterfactuals.[50]

---

[47] Jack Spencer, 'Able to Do the Impossible' (2017) 126 Mind 465, 471.

[48] Mele, 'Agents' Abilities' (n 29). The long discussion of when we are entitled to promise to Φ is something of a red herring.

[49] Maier, 'Agentive Modalities' (n 38).

[50] They are not analytically identical as we can order the distribution of contextual factors according to their output success rate, such that, given certain contextual factors, success is guaranteed, or conversely impossible. Of course, this doesn't entail that they're the *same*: that outcome is an artefact of our ordering. The extent to which they overlap depends on the extent to which the precise context is typical.

(*Possible worlds*. Counterfactuals are typically analysed in terms of possible worlds.[51] CA analysed 'D can Φ' as 'D would Φ if {…}'. On the possible worlds analysis, this means 'D would Φ in possible worlds {w1, w2…}'. The question, then, is in which possible worlds we are interested. We might be interested in the *proportion* of possible worlds in which D Φs, or the *similarity* of the possible worlds to the actual world. This is just another way to precisify the success threshold and the counterfactual, respectively. Hybrid notions include the 'closeness' or 'representativeness' of such possible worlds.)

Once again, however, precisifying the counterfactual, the closeness of the relevant possible worlds, fails to solve the fundamental dilemma. Broad capacity ascriptions are useful but vague and often false. Narrow capacity ascriptions are precise, true, but useless for culpability ascriptions. Return again to the trucker hurtling towards a child. We might describe the counterfactual as follows:

> *Broad*: D is driving towards a child. In this context he can (general) stop, at Φ_{80}.

This capacity ascription is sensible enough. But it is vague, in a way that undermines its appropriateness for culpability ascriptions in many cases, such as:

> *Narrow 1*: D is driving towards a child when a swarm of bees surrounds his cab. In this context he cannot (general) stop, at Φ_{20}.

This context is crucial when ascribing his capacity to stop or not, at least for the purpose of culpability ascriptions. It would be highly misleading to omit such information. Narrow contexts are usually more relevant to the incapacity rules than broad contexts. Aside from infancy and very severe cases of insanity, the vast majority of incapacity rules account for transient or limited incapacities: capacities of understanding *at a particular moment*, of self-control *following threats or provocations*, of belief *given the particular context*. But once we accept that we ought to precisify the context, it's hard to know where to stop. We might want to account for contextual factors such as:

> *Narrow 2*: D is driving towards a child in moderate rain, being on the road for three hours, having eaten four hours ago, half-perceiving a possible moving object forty metres away out of the corner of his eye, without fully drawing his attention as he is worried about his leaking roof {…} In this context he cannot (general) stop, at Φ_{30}.

[51] See the papers in David Lewis, *Philosophical Papers: Volume II* (Oxford 1986) Part IV (Counterfactuals and Time).

The problem with this, of course, is that it simply recapitulates the problem we started out with. To draw the boundaries of capacity ascriptions around so narrow a set of counterfactuals is to eliminate almost all of the factors which would need to be varied to alter the outcome. That leads to a complete retreat to actualism, ie

> *Narrow 3*: D is driving towards a child. In the context of the complete causal chain leading to this moment he cannot stop, at Φ_0.

Precise, but useless for culpability ascriptions.

Let's take stock again. We disambiguated the concept of a capacity by distinguishing particular from general capacities. General capacity ascriptions seemed the better fit for use in culpability ascriptions. But they were vague and misattributed capacities in precisely the sort of cases for which the incapacity rules are designed to account. We tried to remedy those flaws by precisifying the success threshold and counterfactual context for the attribution of general capacities. But this process of precisification collapsed the distinction between general and particular capacities and lead straight to actualism. We are back at square one.

Bullet-biting, again. I started with two responses to the problems with counterfactual capacity ascriptions: actualism and bullet-biting. The failure of disambiguation and precisification leave bullet-biting yet more appealing. Several philosophers concur. Wittgenstein claimed that

> we must be on our guard against thinking that there is some totality of conditions corresponding to the nature of each case (for example, for a person's walking) so that, as it were, he could not but walk if they were all fulfilled.[52]

Alan Millar agrees:

> a statement to the effect that a person has an ability to do something is not shorthand for a longer statement to the effect that the person has the ability to do that thing provided that…, where the gap is filled by specific conditions that together amount to *enabling conditions* for the exercise of the ability. It seems plausible

[52] Ludwig Wittgenstein, *Philosophical Investigations* (1953, Anscombe trs, 3rd edn Basil Blackwell 1967) §183, cited in Millar, *Knowing by Perceiving* (n 38) 127 fn 3. That comment follows the prescient note (at §182) that 'The criteria we accept for … "being able to"…are much more complicated than might appear at first sight…[L]inguistic intercourse that is carried on by… [such] means, is more involved—the role of these words in our language other—than we are tempted to think.'

> that such grasp as we have of enabling conditions for exercising abilities does not in general take the form of a conception of some totality of conditions that could be specified in statements of that sort. [53]

Of course, this bullet-biting—this refusal to analyse capacity ascriptions in any more precise terms—hardly solves the problem of vagueness. Nor are these pronouncements backed by any kind of further argument.

John Maier claims that we might avoid the problems of precisifying modal capacity ascriptions by avoiding modals altogether. He claims that capacity ascriptions can be thought of as *generic* statements instead.[54] But, as far as I can tell, a generic statement remains precisely in need of the sort of precisification discussed above to be workable in practice. That the defendant might *generically* Φ is no guarantee that they can Φ on the relevant occasion for the purposes of capacity/culpability ascriptions.

A final attempt at providing a principled kind of bullet-biting is suggested by Jack Spencer, who suggests 'representative attempts' as the relevant counterfactual. A (hypothetical) attempt counts as representative, claims Spencer, only if D's counterparts are exactly the same, in exactly the same context.[55] But what counts as the same context is left entirely unstated. We can say, ex post facto, that attempting to stop with bees in the cab is unrepresentative, or that being asleep is unrepresentative. But relying on such judgements is simply to retreat to pretheoretical capacity ascriptions, giving up on the attempt to analyse capacities such that the counterfactual analysis can itself tell us what counts as being capable in some contested case.[56] Bullet-biting might achieve intuitive and extensionally plausible capacity ascriptions. But it does so only by abandoning the attempt to analyse capacities altogether. We are left, with Honoré, to be satisfied with capacity ascriptions so long as 'there is no immediate or obvious obstacle' to the agent.[57]

Can we do any better?

---

[53] Millar, *Knowing by Perceiving* (n 38) 127. He reiterates the point at 128ff.

[54] John Maier, 'Ability, modality, and genericity' (2018) 175 Philos Stud 411 §3.

[55] Spencer, 'Able to Do the Impossible' (n 47) 471-473.

[56] Spencer seems to accept this, in part: ibid 471 fn 7, 472 fn 9.

[57] Honoré, 'Can and Can't' (n 35) 158.

Disambiguation, again. We've already disambiguated two senses of capacities and precisified general capacities in two ways (the success rate and the counterfactual). But more disambiguation is possible.

Binary versus scalar capacities. In precisifying the requisite success threshold for general capacity ascriptions, we implicitly accepted that some threshold was required—that capacity ascriptions must be binary. But we could instead prefer a scalar view: one without any threshold. We would no longer ask if Φ_N counts as a 'general' capacity. We would simply say that the agent was Φ_N-able. No more, no less. Of course, that won't do for the incapacity rules. For these require binary judgements. But we can take a scalar view as to the *ontology* of capacities, whilst converting that into a binary capacity ascription for legal purposes. Interestingly, that conversion need not maintain a consistent success threshold. Given that binary ascriptions would be simply artefactual, the choice of threshold might be made on (differing) non-metaphysical grounds.

More context. The relevant success threshold might differ by context even if it is part of the ontology of binary capacity ascriptions. (But especially if it is merely artefactual.) We already noted that the threshold may have to be lowered or raised to account for certain contexts (like being asleep), or even the normative status of the action in question (raising one's hand in class). But it might also differ between particular capacity ascriptions in more subtle ways. It might differ, for instance, depending on the *valence* of the capacity ascription. Some negative capacity ascriptions (incapacity ascriptions) imply that Φing is near impossible, such that we reserve them for (say) Φ_1. ('I could not walk'). At the same time, positive capacity ascriptions regarding the same subject matter might be more generously ascribed. (Perhaps 'I could walk' is permitted 'only' at Φ_{90}). If this is right, there is an asymmetry between the requisite success threshold required for capacity versus incapacity ascriptions.[58]

Act-descriptors. A final ambiguity affects the act-descriptor; the scope of the verb Φ. When playing darts, I can (Φ_{95}) 'hit the dartboard' but cannot (Φ_5) 'hit the bullseye'. There is no correct answer as to whether I can 'make the shot' if that is ambiguous between the two actions. A meaningful answer requires the act-descriptor to be precisified.[59] Likewise, we

[58] I suggest an optimistic asymmetry, but other contexts may favour a pessimistic asymmetry whereby 'could' requires (near) certainty but 'could not' mere probability. Such an asymmetry has been found in folk intuitions regarding ascriptions of personal identity: Kevin Tobia, 'Personal identity and the Phineas Gage effect' (2015) 75 Analysis 396.

[59] This ambiguity results from the ambiguity of actions themselves. As Jennifer Hornsby puts it, 'actions are variously describable particulars': 'On What's Intentionally Done' in Stephen Shute et al

might ask if a keen skier can ski when they are in the hospital with a broken leg. Once again, there is no unitary answer: she can (Φ_{95}) ski when healed, but cannot ski (Φ_{5}) when injured.[60] The act-descriptor might be specified in different ways depending on context via norms of interpretation.[61] Imagine I'm gathering an anthology of Nabokov's mature novels. I would be culpable for failing to read them unless, let us stipulate, I lack the capacity to do so. Can I read Nabokov's mature novels? Nine are in English, one is in Russian.[62] So yes: I can (Φ_{90}). But am I culpable for failing to read the Russian novel? Surely not. The act-descriptor ought to distinguish the two cases. In the criminal law, we do not say that D has the capacity to avoid hitting people (Φ_{99}) and is therefore culpable for so doing if D was sleep-walking when doing the hitting. The two act-descriptors need to be distinguished: hitting when awake and hitting when sleepwalking. Of course, draw the act-description too narrowly and we face all the problems with actualism we noted of contextual precisification.

These ambiguities give rise to various possible specifications for any particular (in)capacity ascription. Take the trucker we started with. Could he have avoided hitting the child? Here are some ways to unpack that question:

A) Would he have avoided the child when driving 80% of the time?
B) Would he have avoided the child when driving in similarly wet conditions 90% of the time?
C) What proportion of the time would he have avoided the child by applying brakes to stop the truck?
D) What proportion of the time would he have avoided the child given his transitory visual impairment induced by flashing lights?

(The law frequently adds: Which of his attributes are relevant to such judgments: age, gender, etc? Those relativisation questions act as prior filters on which attributes may be considered in this process of disambiguation. But the need to disambiguate remains even once we have

---

(eds), *Action and Value in Criminal Law* (Clarendon 1993) 56 fn 3. This follows Donald Davidson's seminal treatment in *Essays on Actions and Events* (2nd edn, Oxford 2001).

[60] Thanks to John Hyman for this example.

[61] This could be via implicature (a la Grice) or impliciture (a la Kent Bach, 'Conversational Impliciture' (1994) 9 Mind and Language 124. (Via Barbara Levenbook, 'Why Ekins's Approach to Statutory Content Fails' (draft)). Angelika Kratzer analyses the semantics of all 'can' statements to suppress '…in view of': 'What 'Must' and 'Can' Must and Can Mean' (1977) 1 Linguistics and Phil 337, reprinted (with revisions) in Angelika Kratzer, *Modals and Conditionals* (Oxford 2013).

[62] Let 1939 be the onset of Nabokov's maturity.

accounted for those relativisations. For we must still ask whether *this* defendant was capable of Φing given all permissible factors.) Questions (A) to (D) might have different answers. They might result in different (in)capacity ascriptions. Perhaps we can intuit which disambiguation, and thus which question, is more appropriate in any particular context. Those philosophers who reject disambiguation and/or precisification rely on this being the case. But we have found no robust analysis for capacity ascriptions, nor a decision procedure for reaching such an answer.

Capacities and causes. Perhaps this shouldn't surprise us. I noted above that capacity ascriptions and causal ascriptions both depend on counterfactuals. Fleshing out the metaphysics of counterfactuals is hard. Philosophers have struggled to analyse causation, just as we have struggled to analyse capacities. Lawyers have struggled too. Michael is driving his camper van quite safely when another car, driven by T, a heroin addict, careens into his path. The vehicles collide, T dies. Did Michael cause T's death?[63] Arthur works successively for company D1 then D2. Both expose him to toxic particulate. Arthur dies, but it cannot be proven under which company he was exposed to the fatal dose of particulate. Did both cause his death?[64] Courts have to determine liability in such cases, by answering yes or no. (They answered no, and yes, respectively). It is commonly suggested that such judgments are morally inflected.[65] Indeed: that judgements of causation for attributing legal liability are morally inflected by definition.[66] Some are happy with such morally inflected causal judgments so long as we are talking about so-called 'legal' causation. But the cases above concerned *factual* causation. Some legal philosophers have suggested that morally inflecting factual causal judgments is inappropriate. It might be acceptable to filter the set of causal attributions for liability ascriptions (so that we don't waste our time focusing on the causal contribution of oxygen to an arson, for instance). But liability ascriptions are supposed at least, the thought goes, to be predicated on some factual cause of a prohibited event.[67] (Naturally, not everyone agrees.) That view presupposes, however, that a good answer—a mechanistic or metaphysical answer—can be given as to whether D factually caused some

[63] *Hughes* [2013] UKSC 56.

[64] *Fairchild v Glenhaven Funeral Services Ltd* [2002] UKHL 22.

[65] Christopher Hitchcock, 'Three Concepts of Causation' (2007) 2 Phil Comp 508, following HLA Hart and AM Honoré, *Causation in the Law* (Oxford 1959).

[66] Hitchcock calls this the 'folk attributive' conception of causation.

[67] AP Simester, 'Causation in (criminal) law' (2017) 133 LQR 416.

outcome. But such good answers might not be available. As Christopher Hitchcock puts it, the scientific conception of causation, as

> described by causal models[,] does not yield causal relations with the grammatical form 'C causes E', where C and E report, events, facts, or states of affairs. The objective causal structure of the world has a more complex grammar.[68]

Perhaps we can update causal attributions in the law to account for a more fine-grained conception of causation. Something like 'partial causation' would seem especially well suited to the toxic particulate case, and suggestions along these lines have been proposed.[69] How is all of this relevant to capacity ascriptions? Well, I noted in chapter 4 that it is capacity ascriptions which drive culpability ascriptions in many cases—not merely aetiological factors. To know if aetiology restrictions were justified, I said, requires that we know the extent of exculpatory incapacitation. But the lesson from both our enquiry above and from the causal literature is that working out what counts as an incapacity is extremely complex. Capacity ascriptions are vague or else admit of many possible disambiguations. We have no reliable or successful means to attributing binary truth functions to ascriptions of causes nor capacities.

3 Capacity and culpability

My analysis of capacity ascriptions has been rather pessimistic. I have not offered any clear way to distinguish *Culpable* from *Excuse*. Without that, there is no clear way to know if the law's aetiology restrictions are justified as providing the best balance of over- and under-inclusivity. I aimed to reach a principled understanding of the incapacity rules. But it seems there is little principle to be found. There are simply many possible factors that might—might—be used to ascribe incapacities, and with them the incapacity rules.

This does allow us, however, to reject some commonly held views about the incapacity rules. For, as with causal attributions, the law (and commentaries) are often premised on the assumption that some principled capacity ascription is in the offing. Consider the following claims by the Law Commission:

---

[68] Hitchcock, 'Three Concepts of Causation' 514.

[69] Michael Moore, 'Moore's Truths About Causation and Responsibility: A Reply to Alexander and Ferzan' (2012) 6 Crim Law and Philos 445, Alex Kaiserman, 'Partial Liability' (2017) 23 Legal Theory 1.

> It is central to [our proposed] defence that the accused was *in*capable of complying with the relevant law. In other words, the defence is only to be available where the accused totally lacked capacity rather than where he or she partially lacked capacity or lacked effective capacity. This limitation is justified in theoretical terms because this is not a defence of reduced responsibility but of no responsibility. It is also justifiable in policy terms because it will help to exclude from the defence those with, for example, a personality disorder which makes it hard but not impossible to control antisocial impulses.[70]
>
> While there are a great many people convicted of offences who have mental health problems and/or learning difficulties, the number who completely lack criminal responsibility as a result is small.[71]

The problem is: how are we to distinguish being 'incapable' from 'partially lacking' a capacity? How are we to be sure that the number who 'completely lack criminal responsibility' is small? These claims are premised on the assumption that incapacity and partial capacity ascriptions are cleanly distinguishable. But that is not true. We can cut capacities at many joints. Being 'impossible' to perform initially sounds like a plausible criterion. But it quickly crumbles under scrutiny. If we mean a general capacity ascription broadly specified (at Φ_{80}), almost no action is impossible. We could narrow the success threshold and raise the contextual precision to conclude that some action was impossible. But doing this generates a large number of false negatives, such that much we don't do would be deemed impossible. In the face of such difficulties, we should be wary of blanket pronouncements about capacity ascriptions. In particular, we ought to be slow to deny claims of incapacities on purely metaphysical grounds. For such grounds may well be spurious. (Likewise, we should be slow to conclude that Michael 'caused' T's death, given the consequences for Michael).

Just as we should be slow categorically to distinguish incapacities from partial capacities, we should be slow too to distinguish incapacities to Φ from Φing being 'merely' *difficult*. Peter Vranas distinguishes two senses in which morality might be too demanding:

> First, morality might be thought to be too demanding in the sense of requiring us to do things that we find very hard to do, things that constitute significant sacrifices.

[70] Law Commission, *Insanity and Automatism* [3.3].

[71] ibid [1.83].

> ...Second, morality might be thought to be too demanding in the sense of requiring us to do things that we literally cannot do, things that go beyond our abilities.[72]

As did the Law Commission, Vranas (and others) take this distinction to meaningfully demarcate two categories. But, again, I doubt this. The difficulty of Φing is simply a function of some success threshold given some counterfactual specification. That is: for Φing to be difficult to some degree simply *is* for one to be incapable of Φing to that degree in some context. Of course, there are many ways for Φing to be difficult: due to duress, circumstances, physical limits, psychological barriers, other incentives, and so on. There may be correspondingly many different ways to disambiguate the claim that Φing is difficult. But all this is equally true of capacity ascriptions. Indeed, the incapacity rules reflect a wide variety of factors that may be unpacked either in terms of rendering defendants less capable or in terms of rendering not offending more difficult. Such claims amount to the same thing.

The incapacity doctrines invoke a wide variety of incapacity ascriptions: incapacities to know the nature of one's act (insanity, diminished responsibility, automatism), to understand right from wrong (insanity, infancy, diminished responsibility), to follow proceedings (fitness to plead), to self-control (loss of control). Relativisations and counterfactual relativisations are predicated upon an unstated and open-ended set of relevant incapacity ascriptions. It would be remarkable if this diversity of incapacity ascriptions ought nevertheless to be understood by reference to an invariant success threshold over an invariant set of counterfactual conditions. We thus face difficult questions of judgment when asked to ascribe incapacities. The considerations canvassed above might help to specify precisely in virtue of what one is ascribing an incapacity. But metaphysics will not come to the rescue to make those judgments for us.

From this, we might retreat to quietism, to bullet-biting, to shrugging and accepting vague capacity ascriptions so long as no obvious defeaters present themselves. But there is another option. Throughout this thesis, I have considered those who object to incapacity rules, or at least to their expansion. A fertile ground for such criticisms was that the law ought not to vary its standards to accommodate frailties and failings—that by doing so it abdicates its standard-setting role. I have consistently resisted that line of critique. If people are less culpable in virtue of their incapacities, then the law ought to account for that lowered culpability. Not at all costs. Not come what may. But insofar as the consequences are not too

[72] Peter Vranas, 'I Ought, Therefore I Can' (2007) 136 Phil Stud 167, 196-197. Reference via Frederick Wilmot-Smith.

bad. Those critics ought to point to some tangible benefit to be derived from denying exculpation to those who are presumptively morally exculpated on the basis of incapacity.[73] But one conclusion to be drawn from this chapter is that the line between incapacity (and exculpation) on the one hand, and capacity (and culpability) on the other is not a definitive matter. The metaphysics of capacities, and the supervening moral implications, remain indeterminate. Perhaps, then, the law could provide some determinacy.[74]

The aetiology requirements, in this light, provide the first pass of an answer. They rule out certain counterfactuals as irrelevant to incapacity ascriptions. But they err on the side of underinclusivity and are designed as heuristics rather than direct determinants for use in capacity ascriptions. That could change. The law could signal which counterfactuals it will deem relevant or irrelevant to incapacity ascriptions. It could set a relevant success threshold. It could set different thresholds for different (elements of) different incapacity rules. It could specify different inferences, judgments, or liabilities to follow from different gradations on a scale of incapacitation. In other words, the law could step in to provide a level of precisification that is absent from the metaphysics of incapacities. Working out precisely *how* it ought to do so, however, is a task for another day.

[73] We ought not to deny exculpation to the *morally exculpated*. But the 'presumptively' morally exculpated? What does that mean? If incapacity ascriptions are either vague or ambiguous, then it would be too easy simply to accept that any alleged or prima facie incapacity truly ought to count as an incapacity, and a fortiori as an incapacity which grounds moral exculpation. Such incapacity ascriptions are on thin ice. But so too are *capacity* ascriptions. If we favour false exculpation over false inculpation, then it is fair to deem those who are allegedly incapably non-culpable presumptively morally exculpated.

[74] Tony Honoré, 'The Dependence of Morality on Law' (1993) 13 OJLS 1.

Bibliography

Aaronson S, 'The Ghost in the Quantum Turing Machine' in Cooper SB and Hodges A (eds), *The Once and Future Turing* (Cambridge 2016)

Alexander L and Moore M, 'Deontological Ethics' in Zalta E (ed), *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition)

Alexander S, 'Contra Caplan on Mental Illness' *Slate Star Codex* 7 October 2015

Aristotle, *Nicomachean Ethics* (Bartlett and Collins trs, Chicago 2011)

—— *Metaphysics Book IX* (Stephen Makin trs, Oxford 2006)

Ashworth A, 'The Elasticity of Mens Rea' in Tapper CFH (ed), *Crime, Proof and Punishment: Essays in the Memory of Sir Rupert Cross* (Butterworth 1981)

—— 'Taking the Consequences' in Shute S, Gardner J, and Horder J (eds), *Action and Value in Criminal Law* (Oxford 1993)

—— 'Ignorance of the Criminal Law, and Duties to Avoid It' (2011) 74 MLR 1

—— and Horder J, *Principles of Criminal Law* (7th edn, Oxford 2013)

Austin J, *Philosophical Papers* (JO Urmson & GJ Warnock eds, Clarendon 1961)

Bach K, 'Conversational Impliciture' (1994) 9 Mind and Language 124

Baron M, 'The Standard of the Reasonable Person' in RA Duff et al (eds), *The Structures of the Criminal Law* (Oxford 2011)

Bear A and Knobe J, 'Normality: Part Descriptive, Part Prescriptive' (2017) 167 Cognition 25

Beard M, *SPQR* (Norton 2015)

Beauchaine T, 'A Brief Taxometrics Primer' (2007) 36(4) J Clin Adolesc Psychol 654

Bennion F, 'Mens rea and defendants below the age of discretion' [2009] Crim LR 757

Benton R, 'Political Expediency and Lying: Kant vs Benjamin Constant' (1982) 43 JHisIdeas 135

Berman M, 'Two Kinds of Retributivism' in Duff RA and Green S (eds), *Philosophical Foundations of Criminal Law* (Oxford 2011)

Bickle J, 'Multiple Realizability' in Zalta E (ed), *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition)

Bierce A, *The Collected Works of Ambrose Bierce: 1909-1912* (Cornell 2009)

Birks P, 'The Concept of a Civil Wrong' in Owen D (ed), *Philosophical Foundations of Tort Law* (Oxford 1995)

Bogg A and Stanton-Ife J, 'Protecting the Vulnerable: Legality, Harm and Theft' (2003) 23 Legal Studies 402

Bourget D and Chalmers D, 'What Do Philosophers Believe?' (2014) 170 Philos Stud 465

Brennan J and Jaworski P, 'Markets without Symbolic Limits' (2015) 125 Ethics 1053

Broome J, *Rationality Through Reasoning* (Wiley Blackwell 2013)

Buse J et al, 'The Modulating Role of Stress in the Onset and Course of Tourette's Syndrome: A Review' (2014) 38(2) Behaviour Modification 184

Byrd S, 'On Getting the Reasonable Person out of the Courtroom' (2005) 2 Ohio State Journal of Criminal Law 571

Cahalan S, *Brain on Fire: My Month of Madness* (Simon and Schuster 2012)

Caplan B, 'The Economics of Szasz: Preferences, Constraints and Mental Illness' (2006) 18 Rat. and Soc. 333

—— *Selfish Reasons to Have More Kids* (Basic 2011)

Caruso G, 'Skepticism About Moral Responsibility' in Zalta E ed, *The Stanford Encyclopedia of Philosophy* (Spring 2018 edn)

Chalmers J and Leverick F, 'Fair Labelling in Criminal Law' (2008) 71 MLR 217

Child JJ and Reed A, 'Automatism is never a defence' (2014) Northern Ireland Legal Quarterly 167

Chituc V et al, 'Blame, not ability, impacts moral 'ought' judgments for impossible actions: toward an empirical refutation of 'ought' implies 'can'' (2016) 150 Cognition 20

Cipriani D, *Children's Rights and the Minimum Age of Criminal Responsibility: A Global Perspective* (Ashgate 2009)

Coffee J, 'Does "Unlawful" mean "Criminal"?: Reflections on the Disappearing Tort/Crime Distinction in American Law' (1991) 71 Boston University Law Review 193

Colvin E, 'Exculpatory Defences in Criminal Law' (1990) 10 OJLS 381

Cornford A, 'The Architecture of Homicide' (2014) 34 OJLS 819

Cowen T, 'The Epistemic Problem Does Not Refute Consequentialism' (2006) 18 Utilitas 383

—— *Stubborn Attachments* (Stripe 2018)

Cullen F, Fisher B, and Applegate B, 'Public Opinion about Punishment and Corrections' (2000) 27 Crime & Just 1

Curwood J, *Hawkins' A Treatise of the Pleas of the Crown Vol I* (8th edition, Sweet, Maxwell and Stevens 1824)

Darwall S, *The Second-Person Standpoint* (Harvard 2006)

Davidson D, *Essays on Actions and Events* (2nd edn, Oxford 2001)

Dempsey MM, 'What We Have Reason to Do: Another View from the Cliff-Top' (2019) 19 JRLS 141

Dennett D, 'Mechanism and Responsibility' in Watson G (ed), *Free Will* (Oxford 1982)

—— *Elbow Room: The Varieties of Free Will Worth Wanting* (Oxford 1984)

—— *Intuition Pumps and Other Tools for Thinking* (WW Norton 2013)

Devlin P, *The Enforcement of Morals* (Oxford 1968)

Dietvorst B and Simonsohn U, 'Intentionally "Biased": People Purposely Use To-Be-Ignored Information, But Can Be Persuaded Not To' (2018) J Exp Psych 1

Dsouza M, 'Criminal Culpability after the Act' (2015) 26 Kings LJ 440

—— 'Intoxication, psychoses, and self-defence: Evaluating *Taj* [2018] EWCA Crim 1743' (2018) Arch Rev

Duarte d'Almeida L, 'O Call Me Not to Justify the Wrong': Criminal Answerability and the Offence/Defence Distinction' (2012) 6 Crim Law and Philos 227

Duff, 'Choice, Character, and Criminal Liability' (1993) 12 Law and Philos 345

— — *Punishment, Communication, and Community* (Oxford 2001)

— — 'Towards a Modest Legal Moralism' (2014) 8 Crim Law and Philos 217

— — 'Cliff-top Predicaments and Morally Blemished Lives' (2019) 19 JRLS 125

Dutton D and Aron A, 'Some Evidence for Heightened Sexual Attraction Under Conditions of High Anxiety' (1974) 30 J Pers Soc Psych 510

Eagle A, 'Twenty-One Arguments Against Propensity Analyses of Probability' (2004) 60 Erkenntnis 371

— — (ed), *Philosophy of Probability: Contemporary Readings* (Routledge 2010)

Edwards J, 'Justice Denied: The Criminal Law and the Ouster of the Courts' (2010) 30 OJLS 725

— — 'Harm Principles' (2014) 20 Legal Theory 253

— — 'Master Principles of Criminalisation' (2016) 7 Jurisprudence 138

Feynman R, *Surely You're Joking, Mr Feynman!* (Bantam 1985)

Foot P, *Virtues and Vices* (California 1978)

Frankfurt H, 'Alternate Possibilities and Moral Responsibility' (1969) 66 The Journal of Philosophy 829

Gardner J, 'Obligations and Outcomes in the Law of Tort' in Peter Cane and John Gardner (eds), *Relating to Responsibility* (Hart 2001)

— — 'The Mysterious Case of the Reasonable Person' (2001) 51 UTLJ 273

— — 'The Wrongdoing that Gets Results' (2004) 18 Philosophical Perspectives 53

— — *Offences and Defences* (Oxford 2007)

— — 'Hart and Feinberg on Responsibility' in Kramer M et al (eds), *The Legacy of HLA Hart* (Oxford 2008)

— — 'What is Tort Law For? Part 1. The Place of Corrective Justice' (2011) 30 Law and Philosophy 1

— — 'Reasons and Abilities: Some Preliminaries' (2013) 58 Am. J. Juris 74

— — 'The Many Faces of the Reasonable Person' (2015) 131 Law Quarterly Review 563

— — 'As Inconclusive as Ever' (2019) 19 JRLS 204

— — and Macklem T, 'Compassion without Respect? Nine Fallacies in R. v. Smith' [2001] Crim LR 623

Goldstein J and Katz J, 'Abolish the "Insanity Defense" – Why Not?' (1963) 72 Yale Law Journal 853

Green L, 'Should Law Improve Morality?' (2013) 7 Criminal Law and Philosophy 473

— — 'The Nature of Limited Government' in Keown J and George R (eds), *Reason, Morality and Law* (OUP 2013)

Hale M, *History of the Pleas of the Crown Vol I* (Nutt and Gosling 1736)

Hájek A, 'Interpretations of Probability' in Zalta E (ed), *The Stanford Encyclopedia of Philosophy* (Winter 2012 Edition)

Hanson R, *The Age of Em* (Oxford 2016)

Harman E, 'Does Moral Ignorance Exculpate?' (2011) 24 Ratio 443

Hart HLA, *Punishment and Responsibility* (2nd edn, Gardner intro, Oxford 2008)

— — and Honoré AM, *Causation in the Law* (Oxford 1959)

Hegel GWF, *Elements of the Philosophy of Right* (1820, Nisbet trs, Wood ed CUP 1991)

Hitchcock C, 'Three Concepts of Causation' (2007) 2 Phil Comp 508

Hodgson J and Tadros V, 'How to Make a Terrorist Out of Nothing' (2009) 72 MLR 984

Honoré T, 'The Dependence of Morality on Law' (1993) 13 OJLS 1

— — *Responsibility and Fault* (Hart 1999)

Horder J, *Excusing Crime* (Oxford 2004)

— — *Homicide and the Politics of Law Reform* (Oxford 2012)

Hornsby J, 'On What's Intentionally Done' in Shute S et al (eds), *Action and Value in Criminal Law* (Clarendon 1993)

Hough M and Roberts J, 'Attitudes to punishment: findings from the British Crime Survey' (1998) Home office research studies

Hume, *An Enquiry Concerning Human Understanding* (1748, PH Nidditch ed, Oxford 1978)

Husak D, 'Rethinking the Act Requirement' (2006) 28 Cardozo L Rev 2437

— — 'Broad Culpability and the Retributivist Dream' (2012) 9 *Ohio St J Crim* 449

— — 'Retributivism in Extremis' (2013) 32 Law and Philosophy 3

— — 'What's Legal about Legal Moralism?' (2017) 54 San Diego L Rev 381

Insel T and Wang P, 'Rethinking Mental Illness' (2010) 303 JAMA 1970

Jahangir Q, Child JJ, and Crombag H, 'Prior fault and contrived criminal defences: coming to the law with clean hands' (2017) Institute of Law Review 1

Kadish S, 'Excusing Crime' (1987) 75 Calif L Rev 257

Kaiserman A, 'Partial Liability' (2017) 23 Legal Theory 1

Kant I, 'On a Supposed Right to Lie from Altruistic Motives' in *Practical Philosophy* (Gregor trs, Cambridge 1996)

Karras RM, *Common Women: Prostitution and Sexuality in Medieval England* (Oxford 1996)

Keating H, 'The "responsibility" of children in the criminal law' (2007) 19 Child Fam LQ 183

Kenny A, *Freewill and Responsibility* (Routledge 1978)

King James Bible

Koch F, 'Skepticism About Special Obligations' (manuscript).

Kolodny N, 'Love as Valuing a Relationship' (2003) 112 Phil Rev 135

— — 'Why Be Rational?' (2005) 114 Mind 509

— — 'Which Relationships Justify Partiality? The Case of Parents and Children' (2010) 38 PPA 37

Korsgaard C, 'The Right to Lie: Kant on Dealing with Evil' (1986) 15 Philosophy and Public Affairs 325

Kratzer A, 'What 'Must' and 'Can' Must and Can Mean' (1977) 1 Linguistics and Phil 337

— — *Modals and Conditionals* (Oxford 2013)

Kugler M et al, 'Differences in Punitiveness Across Three Cultures: A Test of American Exceptionalism in Justice Attitudes' (2013) 103 JCrimLaw and Crimlgy 1071

Lehrer K, 'Cans without Ifs's' (1968) 29 Analysis 29

Leiter B, *Nietzsche on Morality* (Routledge 2002)

—— *Moral Psychology with Nietzsche* (Oxford 2019)

Levenbook B, 'Why Ekins's Approach to Statutory Content Fails' (draft)

Lewis D, *Philosophical Papers: Volume II* (Oxford 1986)

List C, *Why Free Will Is Real* (Harvard 2019)

Locke J, *An Essay Concerning Human Understanding* (1690, Peter Nidditch ed, Oxford 1975)

Loughnan A, *Manifest Madness* (Oxford 2012)

Mackay RD, *Mental Condition Defences in the Criminal Law* (Oxford 1995)

—— 'The abnormality of mind factor in diminished responsibility' [1999] Crim LR 117

—— 'Righting the Wrong? – some observations on the second limb of the M'Naghten Rules' [2009] Crim LR 80

—— and Mitchell B, 'The new diminished responsibility plea in operation: some initial findings' [2017] Crim LR 18

Maier J, 'The Agentive Modalities' (2013) 87 Phil and Phenom Res 1

—— 'Abilities' in Zalta E (ed), *The Stanford Encyclopedia of Philosophy* (Fall 2014 edition)

—— 'Ability, modality, and genericity' (2018) 175 Philos Stud 411

Mandela N, *Long Walk to Freedom* (LBC 1994)

Manwaring J, 'Windle Revisited' [2018] Crim LR 987

—— 'Criminal Law Principles' (draft)

—— 'The Wrongness Limb' (draft)

McKenna M and Coates J, 'Compatibilism' in Zalta E (ed), *The Stanford Encyclopedia of Philosophy* (Winter 2018 Edition)

McMahan J, 'Philosophical Critiques of Effective Altruism' (2016) 73 The Philosopher's Magazine 92

Mele A, 'Agents' Abilities' (2003) 37 Nous 447

— — *Aspects of Agency: Decisions, Abilities, Explanations, and Free Will* (Oxford 2017)

Millar A, *Knowing by Perceiving* (Oxford 2019)

Moore GE, *Philosophical Studies* (Harcourt, Brace & Co 1922)

Moore M, *Law and Psychiatry* (Cambridge 1984)

— — 'Moore's Truths About Causation and Responsibility: A Reply to Alexander and Ferzan' (2012) 6 Crim Law and Philos 445

Morgan R, 'Race and Hispanic Origin of Victims and Offenders, 2012-15' (2017) NCJ 250747: <http://www.bjs.gov/index.cfm?ty=pbdetail&iid=6106>

Morris N, 'Psychiatry and the Dangerous Criminal' (1968) 41 So Cal L Rev 514

— — 'The Criminal Responsibility of the Mentally Ill' (1982) 33 Syracuse L Rev 477

Morse S, 'Crazy Reasons' (1999) 10 J Contemp Leg Iss 189

Mullainathan S and Shafir E, *Scarcity* (Times Books 2013)

Mullis A and Scott A, 'Tilting at Windmills: the Defamation Act 2013' (2014) 78 MLR 87

Nichols S, *Bound: Essays on Free Will and Responsibility* (Oxford 2015)

Nietzsche D, *Daybreak: Thoughts on the Prejudices of Morality* (1881, Clark and Leiter eds, Cambridge 1997)

Norrie A, 'The Coroners and Justice Act 2009: partial defences to murder (1) Loss of Control' [2010] Crim LR 275

Nuotio K, 'On Becoming a Responsible Person' (2005) 2 Ohio St J Crim L 513

Nussbaum M, *Women and Human Development: The Capabilities Approach* (Cambridge 2001)

Ormerod D, *Smith and Hogan's Criminal Law* (13th edn, Oxford 2011)

— — and Laird K, *Smith and Hogan's Criminal Law* (14th edn, Oxford 2015)

— — (ed), *Blackstone's Criminal Practice* (Oxford 2019)

Parfit D, *On What Matters: Volume One* (Oxford 2011)

Pereboom D, 'Determinism *al Dente*' (1995) 29 Nous 21

— — *Living Without Free Will* (Cambridge 2001)

— — *Free Will, Agency, and Meaning in Life* (Oxford 2014)

Platt A and Diamond B, 'The Origins of the "Right and "Wrong" Test of Criminal Responsibility and Its Subsequent Development in the United States: An Historical Survey' (1966) 54 Cal LR 1227

Pinker S, 'The Game of the Name' *New York Times* (New York, 5 April 1994)

Pollock F and Maitland FW, *The History of English Law Before the Time of Edward I* (2nd edn, Cambridge 1898/Liberty Fund 2010)

Popper K, 'The Propensity Interpretation of Probability' (1959) 10 BJPhilSci 25

Poste E (trs), *Institutes of Gaius* (4th edn, EA Whittuck ed, Oxford 1904)

Roberts J, Padfield N, and Harris L (eds), *Current Sentencing Practice* (Sweet & Maxwell 2018)

Posner R, 'An Economic Theory of the Criminal Law' (1985) Columbia Law Review 1193

Quine WVO, *Quiddities* (Harvard 1987)

Raz J, 'Legal Principles and the Limits of Law' (1972) 8 Yale LJ 823

— — 'Promises in Morality and Law' (1982) 95 Harv LR 916

— — *Practical Reason and Norms* (2nd edn, Oxford 1990)

— — *From Normativity to Responsibility* (Oxford 2011)

Ribeiro G, 'The Case for Varying Standards of Proof' (manuscript 2016),

Ritchie H and Roser M, 'Mental Health' (2019) https://ourworldindata.org/mental-health

Ritchie S, *Intelligence: All That Matters* (John Murray 2015)

Roberts J et al, *Penal Populism and Public Opinion* (Oxford 2003)

Robinson P, *The Structure and Function of Criminal Law* (Oxford 1997)

Ronson J, *So You've Been Publicly Shamed* (Picador 2015)

Rosen G, 'Culpability and Ignorance' (2003) 103 PAS 61

— — 'Kleinbart the Oblivious and Other Tales of Ignorance and Responsibility' (2008) 105 Journal of Philosophy 591

Sacks O, *The Man Who Mistook His Wife for a Hat* (Summit 1985)

Sandars TC (ed), *The Institutes of Justinian* (3rd edn, Longmans 1865)

Sarch A, 'Who Cares What You Think? Criminal Culpability and the Irrelevance of Unmanifested Mental States' (2017) 36 Law and Philosophy 707

Sardy M, *The Edge of Every Day* (Pantheon 2019)

Schauer F, *Playing by the Rules* (Oxford 1991)

Schelling, *The Strategy of Conflict* (1960, Harvard 1980)

Schroeder T, Roskies A, and Nichols S, 'Moral Motivation' in Doris J (ed), *The Moral Psychology Handbook* (Oxford 2010)

Simester A, 'Is Strict Liability Always Wrong?' in Simester A (ed), *Appraising Strict Liability* (Oxford 2005)

— — 'Intoxication is Never a Defence' [2009] Crim LR 3

— — 'Wrongs and Reasons' (2009) 72 MLR 648

— — 'A Disintegrated Theory of Culpability' in Baker D and Horder J (eds), *The Sanctity of Life and the Criminal Law: The Legacy of Glanville Williams* (Cambridge 2013)

— — 'Causation in (criminal) law' (2017) 133 LQR 416

— — et al, *Simester and Sullivan's Criminal Law* (4th edn, Hart 2010)

— — et al, *Simester and Sullivan's Criminal Law* (6th edn, Hart 2016)

Sinnott-Armstrong W, "Ought' conversationally implies 'can" (1984) 93 Phil Rev 249

Slobogin C, 'An End to Insanity: Recasting the Role of Mental Disability in Criminal Cases' (2000) 86 Virg LR 1199

Spencer J, 'Able to Do the Impossible' (2017) 126 Mind 465

Stanton-Ife J, 'Strict Liability: Stigma and Regret' (2007) 27 OJLS 151

Stark F, 'Prior Fault' (2014) 73 Cam LJ 8

Stephen JF, 'On what Principles ought the Law to deal with Questions of Responsibility and Mental Competence, in Civil and Criminal cases respectively?' in George Hastings (ed), *Transactions of the National Association for the Promotion of Social Science* (Longman 1865)

— — *Stephen's Digest of the Criminal Law* (1st edn, 1877)

— — *A History of the Criminal Law of England* (1883)

Stern R, 'Does 'Ought' Imply 'Can'? And Did Kant Think It Does?' (2004) 16 Utilitas 42

Strawson G, 'The Impossibility of Moral Responsibility' (1994) 75 Phil Stud 5

Szasz T, *The Myth of Mental Illness: Foundations of a Theory of Personal Conduct* (Perennial 1961)

Tadros V, *Criminal Responsibility* (Oxford 2005)

— — *The Ends of Harm* (Oxford 2011)

— — 'Fair Labelling and Social Solidarity' in Zedner L and Roberts J (eds), *Principles and Values in Criminal Law and Criminal Justice Essays in Honour of Andrew Ashworth* (Oxford 2012)

— — *Wrongs and Crimes* (Oxford 2017)

Tobia K, 'Personal identity and the Phineas Gage effect' (2015) 75 Analysis 396

— — 'How People Judge What is Reasonable' (2018) 70 Ala L Rev 293

Tolmie J, 'Consent to Harmful Assaults the Case for Moving Away from Category Based Decision Making' [2012] Crim LR 656

Valentini L, 'Ideal vs. Non-ideal Theory: A Conceptual Map' (2012) 7 Phil Com 654

Varden H, 'Kant and Lying to the Murderer at the Door... One More Time: Kant's Legal Philosophy and Lies to Murderers and Nazis' (2010) 41 JSocPhil 403

von Hirsch A, *Censure and Sanction* (Oxford 1993)

Vranas P, 'I Ought, Therefore I Can' (2007) 136 Phil Stud 167

Walker N, *Crime and Insanity in England: Volume 1* (Edinburgh 1968)

— — 'The end of an old song?' (1999) 149 NLJ 64

Watson A (ed), *The Digest of Justinian Vols 1-4* (Pennsylvania 1985)

Watts T, Duncan G, and Quan H, 'Revisiting the Marshmallow Test: A Conceptual Replication Investigating Links Between Early Delay of Gratification and Later Outcomes' (2018) 29(7) Psych Sci 1159

Wang EW, *The Collected Schizophrenias* (Graywolf 2019)

Wedgwood R, 'Rational 'Ought' Implies 'Can'" (2013) 23 Philosophical Issues 70

Widerker D and McKenna M (eds), *Moral Responsibility and Alternative Possibilities* (Ashgate 2003)

Williams B and Smart JJC, *Utilitarianism: For and Against* (Cambridge 1973)

Williams G, 'The Definition of Crime' [1955] Current Legal Problems 107

—— 'Recklessness Redefined' (1981) 40 Cam LJ 252

Wilson W, 'The Filtering Role of Crisis in the Constitution of Criminal Excuses' (2004) 17 Can JL Juris 387

Wittgenstein L, *Philosophical Investigations* (1953, Anscombe trs, 3rd edn Basil Blackwell 1967)

Wolf S, 'Sanity and the Metaphysics of Responsibility' in Schoeman F (ed), *Responsibility, Character, and the Emotions: New Essays in Moral Psychology* (Cambridge 1987)

Yaffe G, *The Age of Culpability* (Oxford 2018)

Yeo S, 'The Insanity Defence in the Criminal Law of the Commonwealth of Nations' [2008] Singapore JLStud 241