Organizations as Wrongdoers: from Ontology to Morality

Stephanie Collins

This is the final draft of a book chapter whose definitive version will be published by Oxford University Press.

Chapter 1. The Reality of Organizations

1.1 Overview

Organizations do moral wrong. States pursue unjust wars, businesses avoid tax, charities misdirect funds. Our social, political, and legal responses require guidance. We need to know *what* the thing is we're responding to and *how* we should respond to it. We need a metaphysical and moral theory of wrongful organizations.

This book provides a new such theory, paying particular attention to questions that have been underexplored in existing debates. The underexplored questions include: what's the metaphysical relation between organizations and their members, and how can this help us to make sense of organizations' blameworthiness? Can organizations be blameworthy in all the myriad ways 'blameworthy' is understood in contemporary moral theory? And what about feelings of guilt, remorse, and shame—can organizations experience these feelings in a phenomenological sense (not just their 'functional equivalents')—and why does this matter for organizations' blameworthiness? And if organizations are wrongdoers 'in their own right,' then what does this imply for members? How and why are members implicated in organizations' wrongs, and how should reparative costs be apportioned among members?

This book will answer these and other questions. In doing so, I aim to bridge a divide. On one side of the chasm, recent work in social ontology has aimed to account for the metaphysics of organizations. But this work has not considered how the metaphysics can address the distinctive moral issues raised by organizations' wrongdoing. On the other side of the chasm, recent work has taken groups' moral blameworthiness seriously, but without fully establishing the underlying metaphysics and exploring how that metaphysics can inform our moral conclusions.[1] In short: the metaphysical and moral treatment of organizations have become detached in the literature. This book aims to attach them and to demonstrate why that attachment matters.

The book's primary metaphysical thesis is that members are *material parts* of organizations—much as the engine, wheels, and so on are the material parts of a motorcycle. I argue for this thesis in Part I (Chapters One, Two, and Three). In Part II (Chapters Four and

[1] For example, List and Pettit 2011 and Tollefsen 2015 address group responsibility with a focus on the underlying philosophy of mind; Isaacs 2011 with an emphasis on philosophy of action.

Five), I demonstrate that this metaphysical thesis yields important moral conclusions: organizations can be blameworthy for a wide range of actions, attitudes, and character traits (including those of members qua members) and organizations can literally feel guilt (when members feel guilt qua members). In Part III (Chapters Six and Seven), I explore the implications of organizations' wrongdoing for members. I catalogue the ways in which members-as-material-parts can be implicated in organizations' wrongdoing. As we will see, not all members are equally implicated in each wrongdoing of each organization. There are some organizational wrongdoings in which only a few members are implicated, despite the fact that all members are material parts of the organization that has done wrong. I argue that the costs of organizations' wrongdoing should be distributed to members in proportion to members' level of implication in the wrongdoing, rather than being distributed equally on the basis of membership-as-parthood alone.

In these ways, I aim to connect the metaphysics and morality of wrongdoing organizations. Of course, not all readers will be interested in this connection. Those who are interested only in the metaphysics might stop reading after Chapter Three. I hope that Chapters Two and Three will appeal to anyone interested in the metaphysics of ordinary objects—not only those interested in the metaphysics of *social* objects. Meanwhile, those who are interested only in the moral and political implications can skip Chapters Two and Three, without much loss of understanding (though such readers will, I think, be interested in Sections 3.4 and 3.5). Such readers are asked simply to grant me the assumption that members are material parts of organizations. I hope that Part III (Chapters Six and Seven) will engage anyone with an interest in historical injustice and reparations, regardless of their interest in metaphysics. Still, the book's arguments are best understood as forming a coherent whole: the metaphysical picture is an important premise in the moral arguments, while the moral arguments explain why the metaphysical picture matters.

To settle ideas and begin to demonstrate the practical importance of this inquiry, let me lay out two examples.

Australian Banking Royal Commission

In 2019, Australia received the final report of its Royal Commission into Misconduct in the Banking, Superannuation and Financial Services Industry. In Australia, Royal Commissions engage in inquiry and fact-finding. They're less formal than a court, but they have the power to refer entities to courts for prosecution. Perhaps most importantly, their policy recommendations are taken seriously by government and the media.

The Banking Royal Commission (as I'll call it) was set up to inquire whether there was "misconduct, or conduct falling short of community standards and expectations, in Australian financial services entities." If so, the Commission was asked to report on whether that conduct was "attributable to the particular culture and governance practices of a financial services entity or broader cultural or governance practices in the relevant industry or relevant subsector" and whether the conduct "result[ed] from other practices, including risk management, recruitment and remuneration practices of a financial services entity, or in the relevant industry or relevant subsector." (2018, 2)

The Commission revealed that Australia's biggest banks had engaged in widespread and pervasive dishonest practices. These included charging fees when no service had been provided (including to deceased customers), lying to customers, forging customers' signatures, impersonating customers, falsely witnessing documents, transferring customers' funds to advisors' personal bank accounts, underpaying interest on term deposits, and on and on. The Commission's report attacked the "culture" of "dishonesty and greed" in Australia's largest banking and finance corporations (Royal Commission 2018, 73; 2019, 138). The Commission's final report asserted that the corporations themselves were culpable for the misconduct the Commission uncovered (2019, 4).

Yet the Commission didn't explicitly endorse the idea that the banks were blameworthy in a manner 'irreducible' to the blameworthiness of the individuals involved. Instead, the Commission simply raised the question: "[s]hould there be more focus on criminal proceedings against [banks and financial institutions] rather than individual advisors?" (Royal Commission 2018, 158) It left answering this question to the politicians and bureaucrats who were the primary audience of its report. In this, the Commission arguably submitted to a long tradition in Australia of viewing corporations as 'legal fictions,' with individuals (such as financial advisors) being the ones who are chased by the legal system. The Commission left it up to regulators whether this tradition should be overthrown.

You might think the Commission's question is a legal one, not a philosophical one. But Australian law gives a philosophically contentious answer. Australia's Criminal Code "applies to bodies corporate in the same way as it applies to individuals"—yet it allows for prosecution of a corporation only when "an offence is committed *by an employee, agent or officer* … within the actual or apparent scope of his or her employment, or … authority." (*Criminal Code Act 1995* (Cth), Part 2.5, Sec. 12.1, emphasis added.) Crimes can also be attributed to a "corporate culture," understood as "an attitude, policy, rule, course of conduct or practice existing within the body corporate generally or in the part of the body corporate in which the relevant activities

[of employees, agents, or officers] take place." (*Criminal Code Act 1995* (Cth), Part 2.5, Sec. 12.4) But even within this 'corporate culture' provision, identifiable wrongful activities *of individuals* are what's attributed to the corporate culture. And the corporate culture provision has rarely been used, so there is little clarity over how to interpret and apply it (Hill 2003; Adams et al 2017). The upshot: legally, in Australia at least, corporate criminal wrongdoings generally exist only when individual criminal wrongdoings exist.

In recent years, several philosophers have provided pictures of collective wrongdoing that allow us to argue that this legal orthodoxy is wrong-headed (e.g., French 1984; Pettit 2007; Isaacs 2022; Tollefsen 2015). However, as I aim to show, our current philosophical picture of corporate wrongdoing is incomplete. Therefore, our potential to provide a full and defensible answer to the Commission's question is also incomplete. This book will fill the gaps, providing the theoretical tools for us to answer the Commission's question of whether it is desirable—or even defensible, or even possible—to hold organizations themselves culpable.

What's true of legal wrongdoing is also true of legal metaphysics: Australian law currently holds an individualist picture. The above legal approach to corporate criminal wrongdoing reflects the fact that much of Australia's corporate law seems to employ a shareholder-based metaphysics of corporations, in which the corporation is identified with its shareholders (Hill 2003; this view has recently received philosophical defence in Ludwig 2017a)—or, only slightly less reductively, with a "nexus of contracts" that includes employees, suppliers, customers, and so on (Jensen and Meckling 1976).

The nexus-of-contracts model arguably encourages the reduction of the corporation and its wrongdoing to individuals and their wrongdoing: on this model, if we can find out what those in the nexus have done, then we will have found out what the organization has done. But organizational wrongdoing is both narrower and broader than this. It's narrower insofar as many in the contractual nexus—such as suppliers and customers—should not have their actions attributable to the corporation. And it's broader insofar as wrongs sometimes seem attributable to corporations' *cultures* (as the Banking Royal Commission claimed). It's unclear where to locate a 'culture' within a nexus-of-contracts metaphysics. All of this suggests that philosophy can help to answer the Commission's question—in both its moral and metaphysical dimensions.

The Polluter Pays Principle

A second example is the role of the Polluter Pays Principle in distributing the costs of mitigating the devastating effects of climate change—and, in particular, in implementing the 2015 Paris

Agreement under the United Nations Framework Convention on Climate Change. The Polluter Pays Principle asserts that industrialised states should set high emissions reduction targets under the Paris Agreement, because these states are culpable for wrongfully high historical emissions. Disagreement over the truth of the Polluter Pays Principle (among many other things) has prevented states from setting sufficiently ambitious emissions reduction targets under the Paris Agreement. Predictably, developing states argue in favour of the principle while developed state resist it (Stavins 2018).

If we are to vindicate the Polluter Pays Principle, we need a conception of states as unitary actors, who nonetheless relate to their citizens in such a way that (i) citizens' polluting actions are attributable to states and (ii) costs attributed to polluting states can justly be passed on to a wide diversity citizens—including passed on to citizens who are not themselves wrongful polluters. Parts I and II of this book vindicate point (i), while Part III vindicates point (ii).

As with Australia's Banking Royal Commission, the law gets us only so far in sorting out these issues. *Brownlie's Principles of Public International Law* (Crawford 2012) is the canon on states' legal treatment. Here, states are metaphysically conceptualised as unitary actors rather than a 'nexus of contracts' or a collection of individuals. But in the place of the nexus of contracts, there's an empty hole: states' relation to individuals is left entirely untheorized. It's as if states are completely separate from their individual constituents—whether these constituents are conceptualised so as to include ordinary citizens, or not. Part I of this book aims to fill the hole.

And needless to say, the force of international law has not been used against states who are high emitters. None of them have been condemned by the United Nations Security Council or referred to the International Court of Justice, for example. Philosophy can help us to work out whether that would be morally appropriate. Although Chapters Two and Three will argue that citizens are the *material parts* of their states, Chapter Four will demonstrate that many of the moral failings of organizations—including states—cannot be located in the wrongful actions of citizens. Organizations' moral failings often lie in organizations' *evaluative attitudes* or *character traits*, where these organizational attitudes and traits are not straightforwardly understandable in terms of the wrongful evaluative attitudes or wrongful character traits of members. This analysis reveals the shortcomings of existing treatments of organizations' wrongdoing, which have tended to focus on organizations' blameworthiness for their irreducibly group-level *choices* or *decisions*. If we can hold organizations blameworthy for their evaluative attitudes and their character traits (not just their choices or decisions), we potentially get many more environmentally negligent organizations on the hook.

These two examples will recur throughout the book's discussion, helping to illustrate and problematise various claims that I'll consider and defend. But this is a book of philosophy. My arguments will remain theoretical and general, with the details sketched primarily via these two examples. To introduce the book's philosophical approach, the rest of this introductory chapter turns to theoretical issues. I will first outline how I conceptualise organizations. I'll then briefly explain why we should be realists about them. Finally, I will give an overview of the arguments to come in the rest of the book.

1.2 Organizations' Attributes

In outlining the Banking Royal Commission and the Polluter Pays Principle, I've made an assumption that may surprise some readers. I've assumed states are organizations. Often, when we hear 'organization,' we think of large corporations and charities, but not states.[2] My usage is broader. As I'll use the term, an 'organization' is *a collective agent that involves a large number of people who realise a structure that coordinates divided labour via rules and hierarchical command relations, guided by a collective decision-making procedure.*

This definition is intended to be ecumenical between many different characterisations found across the academic literature. It follows a long tradition in sociology, starting with Max Weber's characterisation of bureaucracies as involving rules and regulations, a division of labour and responsibility, and hierarchical authority structures (1968, vol. I, 223ff; 1968, vol. III, 956ff.). More recently, sociologists have given similar analyses of institutions, such as Rom Harre's view of an institution as "an interlocking double-structure of persons-as-role-holders or office-bearers and the like, and of social practices involving both expressive and practical aims and outcomes" (1979, 98) or Jonathan Turner's definition of an institution as "a complex of positions, roles, norms and values lodged in particular types of social structures and organizing relatively stable patterns of human activity" (1997, 6). My use of 'organization' rather than 'bureaucracy' or 'institution' reflects recent philosophical usage, in which 'bureaucracy' has fallen out of the lexicon, while 'institution' is often used to capture a wider range of phenomena—including institutions such as marriage or promise-keeping, which are not 'organizations' in any sense.

---

[2] Though I'm certainly not the first to argue that states are collective moral agents (see e.g. Erskine 2001; Erskine 2003; Wendt 2004).

My characterisation of organizations also roughly matches recent management theory, particularly Geoffrey M. Hodgson's distinction between social structures, institutions, and organizations. According to Hodgson, "[s]*ocial structures* include all sets of social relations, including the episodic and those without rules, as well as social institutions. *Institutions* are systems of established and embedded social rules that structure social interactions. *Organizations* are special institutions that involve (a) criteria to establish their boundaries and to distinguish their members from non-members, (b) principles of sovereignty concerning who is in charge and (c) chains of command delineating responsibilities within the organization." (Hodgson 2007)

In addition to sociology and management theory, my characterisation resonates with related work in philosophy. Thus my 'organizations' include Peter French's 'conglomerates,' which have (i) internal organization or decision procedures for choosing courses of action, (ii) enforceable codes of conduct for members, and (iii) members who replaceably occupy roles that provide them with power over other members (French 1984, 13ff.). My view likewise dovetails with Raimo Tuomela's view of a group agent as "an interactive social system … that consists of interrelated individuals such that this system is, through them, capable of producing uniform actions and outcomes" (2013, 21).

Because my definition encompasses the above definitions, I hope that my metaphysical and moral claims will appeal to a range of different theorists, who might subscribe to different of the above-cited definitions and characterisations of organizations. That is to say, my aim here is not to *improve* on any of the definitions quoted above from other theorists. Rather, I aim for my definition to be consistent with—and neutral between—all the above definitions. That said, my definition raises a question: what makes organizations (as I've characterised them) worthy of a unitary analysis as wrongdoers, despite the vast differences between various types of organization—states, for-profits, and non-profits, for example? After all, many philosophers differentiate organizations according to those organizations' ends (e.g. French 1984; Rovane 1998; Miller 2010). One might think that organizations with different ends should be conceptualised differently as wrongdoers.

On the contrary, the various components of my definition demonstrate that organizations warrant treatment as a class. First, consider organizations' large size. By 'large,' I mean 'large enough that not all members are known to one another personally.' The numerical cut-off here will differ for different organizations, depending on how channels of communication are set up. Definitionally, all organizations are large enough that members cannot have pairwise in-depth conversations about the organization's plans and policies (at least, not consistently with

having the time to ensure those plans and policies are enacted). This is significant, because it means there is extremely unlikely to be perfect uniformity amongst members' reasons for supporting or opposing particular plans or policies, or even their interpretations of the content of plans and policies. The members cannot gather around a table to reach a consensus view on the whys and hows of organizational practices and policies. This mandates heterogeneity in how we treat members of wrongdoing organizations, as Part III of this book will explain. Such heterogeneity is common to organizations, while not always being true of smaller, more collaborative, collective agents. The lack of cohesion entailed by organizations' large size also means that organizations' members cannot simply be picked out by the fact that they all share identical goals, visions, or values. A more complicated account of membership is required, as Chapters Two and Three will explain.

Second, the members occupy places in a structure. A structure is a collection of roles that stand in relations. In an organization, the roles are jobs or tasks and the relations are usually reporting and delegation lines. The roles and their relations can be highly diverse. Both the roles and the relations are representable in an organization chart (that depicts the roles as 'nodes' and the relations as 'edges').

Organizations are composed of people who 'realise a structure'—that is, occupy the roles in the structure. A structure is an abstract representation of various roles and of the relations between those roles. As an analogy, we can think of a recipe. A recipe is an abstract representation of various roles (the onion role, the garlic role, and so on) and of how the occupants of those roles must be related to one another in order to realise a certain dish. For the dish to be realised, we must find particular role-occupants (particular onions, particular bulbs of garlic, and so on) and we must make it the case that those role-occupants are related in such a way as to follow the recipe.[3]

Following Katherine Ritchie, saying that an organization is composed of people who realise a structure allows us to say that an organization *is* a realised structure, where a structure is 'realised' when enough people occupy the structure's roles. (How many is 'enough'? That will be dictated by the organization's decision-making procedure—something I will characterise shortly.) This idea of structure gives us an explanation of how an organization is

[3] The recipe analogy comes from Kathrin Koslicki's (2008) discussion of ordinary objects.

both 'one' and 'many.' It is 'one' because it is a realised structure—a single thing. It is 'many' because it is the many individuals who occupy its roles.[4]

Chapters Two and Three will say much more about the metaphysics here. Specifically, I will argue that members are *physical parts* of an organization, and that an organization is *located* wherever its members are. This will incorporate organizations into a naturalist, materialist view of the world. For now, the important point is this: since all organizations are realised structures, the just-mentioned metaphysical views can apply in the same way to different types of organizations (states, for-profits, non-profits, and so on). This vindicates the project of developing a unified theory of organizations as wrongdoers.

Third, an organization has a division of labour. Members don't know all the details of what all the others are doing vis-à-vis the organization's plans and policies. This complicates the extent to which each can be held blameworthy for what the others (fail to) do. Diminished individual blameworthiness has two implications. First, it generates an imperative that we have a sound theory of distinctively organization-level wrong, since often we won't find individual-level wrongs that are proportionate to an organization-level wrong. Part II of this book (Chapters Four and Five) theorises distinctively organization-level wrongs, including distinctively organization-level guilt. Second, diminished individual blameworthiness implies that we need to take care when spelling out the implications of organizations' wrongdoings for members, since those implications will not be as uniform as they would be in a group with less informational asymmetry between members. Those implications for members are spelled out in Part III (Chapters Six and Seven).

Fourth, an organization has rules and hierarchical command relations via which it pursues its plans and policies. These matter because—similar to the informational asymmetries that result from the division of labour—they sometimes attenuate individuals' blameworthiness. If an individual has perfectly good reason to trust the system, then she may be perfectly within reason to obey the rules and/or her superiors. Yet if the rules themselves, or the command

[4] See Ritchie (2013). Harris (2020, 360-361) objects to Ritchie's particular version of the organizations-as-structure view: structures do not allow for changes in members, roles, or relations; clearly, organizations can change members, roles, and relations while remaining the same organization. We can respond to this by saying that two realised structures, which exist at different times, are the same organization just in case the later structure is descended in the right way from the earlier structure.

relations, are morally wrong, then this can lead to the organization doing wrong. This phenomenon is common across a wide range of organizations, even if the precise details vary.

Worse, because the rules and command relations are themselves designed through a process with a division of labour (including through legislation and regulation that is imposed upon organizations from the outside), there may be no individual who is blameworthy for the fact that the rules and command relations are as they are. This relates to Anthony Downs' 'Law of Imperfect Control,' according to which no one individual can fully control a large organization—as well as his 'Law of Diminishing Control,' according to which the more effort at control is made at the top, the more subordinates will try to evade or counteract such control (Downs 1967, 143-150). Thus, rules and hierarchical command relationships can relieve *both* superiors *and* subordinates from blame. Of course, this is not to say individuals within organizations are never blameworthy: Chapter Six will spell out many ways in which individuals can be implicated in organizations' wrongdoing, and Chapter Seven will justify individuals bearing costs to repair their organizations' wrong. Again, though, these phenomena are common across a wide range of organizations, vindicating the project of developing a unified theory of organizational wrongdoing.

Thus, organizations are distinctive by virtue of being composed of a large number of people who are occupy places in a structure that coordinates divided labour via rules and hierarchical command relations. This makes them worthy of a unified analysis. And there are good reasons to want a unified analysis. By aiming at unity, our theory is forced to fit with a wider range of case studies and examples. All else being equal, this makes our theory better. That is: if we developed a theory of wrongdoing for (say) corporations but not states, then we'd need to develop a separate account of states as wrongdoers. Our over-arching theory would therefore be less unified and less simple. If we can develop a theory that applies to all organizations in a unified way (as I suggest we can), then this is preferable to going back to the drawing board for each and every different type of organization.

At the same time, though, it's worth theorising about organizations specifically, rather than about collective agents more generally. This is because organizations raise their own problems, as just summarised. While the theory developed in this book might apply to collective agents more generally, I want to make sure we get organizations right.

What are 'collective agents more generally'? I said earlier that organizations are *collective agents* that are *guided by a collective decision-making procedure*. In my view, all collective agents are guided by a collective decision-making procedure (Collins 2019, ch. 6). Specifically, a collective agent is composed of agents united under a rationally operated group-

level decision-making procedure that can attend to moral considerations. Thus, each collective agent has its own decision-making procedure. Importantly for organizations specifically, a collective's decision-making procedure includes the informal, tacit, and vague procedures that we might subsume under the notion of an organization's 'culture' (Weick and Sutcliffe 2001, ch 3; Herzog 2018). Why include 'culture' within a collective agent's 'procedure'? Because the norms, practices, and traditions that constitute a culture often hugely constrain the real-world instantiation of any abstractly defined 'formal procedure.' Think, for instance, of a committee whose formal procedure is 'egalitarian discussion amongst committee members.' When this formal procedure is actually instantiated, culture will inevitably have a strong influence. For example, tradition and norms might informally decree that older male members speak first, or for longer, or with more authority. Such cultural aspects of an actually-instantiated procedure will often be crucial for understanding the procedure's operation and outputs. Chapter Three will say more about organizational membership and how members relate to an organization's decision-making procedure.

My account of collective agency produces a broad church. Two friends deciding where to go for lunch are a collective agent, just as much as BP, France, or Oxfam. Thus, many collective agents are far more unstructured, undivided, discretionary, small, and egalitarian than organizations. For the purposes of analysing collective agents' *wrongdoing*, it is useful to study organizations specifically, rather than the broad church of collective agents more generally (as was done by, for example, Isaacs 2011; List and Pettit 2011; Tollefsen 2015). This is because of the factors mentioned earlier: organizations are (i) composed of many people, (ii) who realise a structure, (iii) with a division of labour, (iv) governed by rules and hierarchical command relations. These factors produce the complications sketched above, which will be discussed as the book proceeds.

1.3 Realism about Organizations

So far, we have two examples where organizations' wrongdoing is at stake but contestable. And we have a conception of organizations that renders them a distinct category. There is one more preliminary to establish in this chapter: realism about organizations. This is the view that organizations "are entities over which we quantify in the set of our best descriptions and explanations of the social world." (Sheehy 2006, 132)

As with any philosophical position, realism about organizations is contested—not just in philosophy, but also in organization theory, business ethics, sociology, and economics.[5] It is not hard to see why it is contested. We don't seem to bump into organizations in the street. We might bump into their buildings or billboards, but these things are+ buildings or billboards—they are not, themselves, organizations. It is not clear where organizations *are*. (I will defend a view of where they are in Chapters Two and Three.) And if it's not clear where they are, then it's not clear how they could possibly be part of the causal pathways of the natural world. Those causal pathways require a spatio-temporal location, or so it seems. Furthermore, when we look out into the social and political world, what we *do* see is individual humans. Our society and politics seems to be made up of humans. While we might sometimes refer to groups of humans, it is tempting to view this as 'mere shorthand' for humans themselves.

Despite these criticisms, realism is widely enough accepted—and the reasons for it have been widely enough circulated—that my defence here will be brief, schematic, and familiar to many readers.[6] This truncated defence will enable the remaining chapters to forge new ground. There are three broad reasons for realism about organizations: their multiple realisability, their inward and outward causal-explanatory power, and (what I'll call) their functionalist-interpretivist agency.

---

[5] Recent sceptics include Heugens, Kaptein, and van Oosterhout 2008, Rönnegard 2015, and Ludwig 2017a—tracing back to Popper 1945, Hayek 1948, Weber 1968, Weick 1979, and Elster 1989.

[6] See, e.g, Goodpaster and Mathews 1982, Ruben 1985, Pettit 1996, Hatch 1997, Reed 2001, Fairclough 2005, Hess 2013, Thomasson 2019. In her influential overview of organization theory, Hatch (1997) divides organization theory into 'modern,' 'symbolic,' and 'postmodern' perspectives. In the sense I'm using here, all three perspectives are 'realist.' The modern perspective analyses organizations as agents of rational control; the symbolic perspective views organizations as cites of cultural meaning and symbolism; and the postmodern perspective views organizations as ideologically-driven forces of hegemonic power. My arguments can be wielded from any of these perspectives. More generally, the metaphysical and moral issues addressed in this book cross-cut debates in organization theory about organizations' interests, values, and meanings.

Multiple Realisability

There are multiple ways organization-level facts, properties, events, and explanations—as well as organizations themselves, as objects—can be realised by individual-level facts, properties, events, explanations, and objects. (The reader might doubt that organizations are objects; I will render this plausible in Chapters Two and Three.) Consider an *event*: the event of Australia permitting a large new coalmine to be built. We could instead have considered *the fact* that Australia has permitted the mine, or Australia's *property* of having permitted the mine, or an *explanation* of the mine that includes Australia's having permitted it, or the *entity* that is Australia—all of these are 'multiply realisable' in relevant individual-level phenomena. By 'multiply realisable,' I just mean that individual-level phenomena could have been a variety of different ways, consistent with the organization-level phenomenon remaining exactly the same. I focus on phenomena that are events, just to settle ideas. (For discussion of the differences between the multiple realisability of social-level properties, facts, events, and so on, see List and Spiekermann 2013.)

The individual-level events that constitute the Australia-level event most prominently include actions on the part of state bureaucrats. And each of these individual-level events could have gone various ways: individuals could work more or less reluctantly, callously, carefully, obediently, etc. Each of these possible realisations is 'Australia permitting the mine,' just so long as two conditions are met.

First, in each realisation, the individual actions are performed by members within the constraints of, and because of, their role in Australia; or, as I will equivalently say, the actions are performed by members *while they are performing their role*. This is what makes the event one of *Australia* permitting the mine, rather than one of some private individuals abusing Australia's procedures to produce a permission for the mine. Second, the permission of the mine results from each realisation. This is what makes the event one of Australia *permitting the mine*, rather than merely *trying* to permit the mine.

If these two conditions are met, then each realisation is a realisation of *Australia's permitting the mine*. The realisations are different from each other, since they each involve individuals performing different actions. Thus, the realisations are not identical to each other. It's also not true that all the realisations are identical to the permitting of the mine—if they were, and if we accept transitivity of identity, then the realisations would also have to be identical to one another (which we've already ruled out, because they are discernible from one another). We also cannot arbitrarily choose just one realisation, and say that the permitting of the mine is identical to *that* realisation but not any of the others: if we did that, then the

permitting of the mine would not occur if one of the other realisations happened. So, we should say that the permitting of the mine is something distinct from, but that unites, the distinct realisations (Jackson and Pettit 1992; List and Menzies 2009). This is just to say that we should be realists about the organization-level event.

Importantly, if the multiple realisability of some organization-level phenomenon is to render it distinct from its individual-level realisers, then the different potential individual-level realisers must be of different types, not just different tokens of the same type (Shapiro 2000; Couch 2004). To see this, consider the analogous case in philosophy of mind. Suppose my pain is always realised by C-fibres firing in my brain, though different C-fibres fire on different pain-occasions (because C-fibres degenerate and are replaced by new ones). Then, pain is equivalent to C-fibres firing. For pain to be non-equivalent to (i.e., distinct from) its neural realisers, there must be different *types* of potential neural realisers. This holds in our example: Australia's act of permitting the mine could have been secured by legislation, or by a judicial ruling, or by the run-of-the-mill implementation of some existing policy, or in some other way. These are all different *types* of realisers of the organization-level event.

Other philosophers have similarly used multiple realisability to argue for organizations' existence, but those arguments have tended to focus on how an organization manifests itself across time (e.g. List and Pettit 2011; List and Spiekermann 2013). Such cross-temporal arguments says that an organization persists through time even while its members alter. For example, an organization persists even though some members leave while other members join. Such persistence across changes in membership provides reason to view the organization as a metaphysically distinct entity. But such cross-temporal change is not the only relevant kind of multiple realisability.

Instead of tracking the organization across time, we can track the organization across different possible worlds: even if members never change in the actual world, we should be realists about the organization if it's true that the organization *would* persist *were* members to change. The counterfactual (rather than cross-temporal) version of multiple realisability is worthwhile for a few reasons. First, it enables realism about a broader pool of organizations than the cross-temporal version of the argument. Second (and more importantly), it enables us to assess just how *robust* the organizational entity is. We can ask across how many changes in membership (or across what kinds of changes in membership) the organization would remain. Of course, any survival across changes in membership gives us reason to posit the organization. But a more robust entity—an entity that would survive a higher or more diverse range of

changes to its membership—is likely to have more explanatory power, and so more reason to be posited. This brings us to the second reason for realism about organizations.

Inward and Outward Causal-Explanatory Power

Above, I mentioned that one might question organizations' explanatory power, because one might question where they are placed in the natural world. I will address that issue at length in Chapters Two and Three. Bracketing that issue for now, we can see that organizations have causal-explanatory power in two directions. They have *inward* causal-explanatory power insofar as organizational facts, properties, events, and entities can causally explain facts, properties, events, and entities internal to the organization. And organizations have *outward* causal-explanatory power insofar as organizational facts, properties, events, and entities can causally explain facts, properties, events, and entities external to the organization. Inward causal-explanatory power implies that (an event, fact, property, etc that involves) the organization sometimes causes (an event, fact, property, etc that involves) its *components* being certain ways. Outward causal-explanatory power occurs when (an event, fact, property, etc that involves) an organization causally explains (an event, fact, property, etc that involves) phenomena *external* to the organizations. Outward causal-explanatory power demonstrates that organization-level explanations are different from an aggregation of explanations that are about the organization's components.

To intuitively see inward causal-explanatory power, consider again the case of Australia permitting a new coal mine. To adequately explain or describe the members' actions that constitute the organization's action—to capture members' motivations, constraints, goals, etc—we need to refer to the organization's distinct features (e.g. judgments, goals, intentions, procedures). We can imagine a bureaucrat who is entirely opposed to permitting the mine—who even secretly campaigned against it on weekends—yet who does her role in implementing Australia's decision. To fully explain and motivate this individual's actions, we need to refer not just to facts about her, but also to facts about Australia—specifically, that Australia permitted the mine. Her action is, in a sense, not her own (at least, it's not *just* her own). The locus of agency behind her action is (also) Australia's. This is the same sense in which the wrongdoings uncovered by the Banking Royal Commissions were problems *with the banks*, not just with individuals within the banks, and provides a rational for why the Polluter Pays Principle targets historically emitting *states*, not just collectives or individuals governed by those states. (For similar points on inward explanatory power, see Elder-Vass 2007, 32-34.)

To intuitively see *outward* causal-explanatory power, consider that organizations can cause effects their members could not cause, were those members to lack the properties in virtue of which they realise the organization's structure. The coal mine—situated on public land in Australia—could not be permitted by one person acting alone (not even the Prime Minister, without violating due process and therefore not acting *as* Prime Minister). The coal mine couldn't even be permitted by many Australian people together, if they were not acting within the structure governed by the organization-level decision-making procedure of Australia. Of course, the mine might end up on the land by sheer fluke, without any action by the Australian state, but then we would not describe the mine as having been *permitted* by anyone. The 'permitting' in that kind of case is not an intentional action, but rather an accident or omission. By contrast, if the mine is permitted by an edict produced by Australia's organization-level decision-making procedure, then the permission is reliably and intentionally controlled; therefore, the permission is an action (Strand 2013). Yet no individual can successfully produce the permission on their own: the action must be attributed to the state.

To these intuitive demonstrations, we can add that some of the leading theories of causal explanation are amenable to including organizations in causal explanations. This is most obviously true on counterfactual theories of causal explanation (see, e.g., List and Menzies 2009; List and Spiekermann 2013). According to these theories, an outcome is explained by the facts that hold across all (or most) of the possible counterfactual worlds where that outcome obtains. Consider Australia's permitting of the coal mine. This outcome doesn't co-vary with the specifics of particular individuals: the individuals are more or less replaceable when it comes to securing the outcome. Instead, it co-varies with organization-level facts (events, properties, entities) like 'Australia's belief that coal creates jobs and boosts GDP.' Such organization-level phenomena also feature in causal explanations that use a probabilistic theory (on which causes make effects more likely) or a robustness theory (on which causes make effects more secure) (See similarly List and Spiekermann 2013. I say more about social-level phenomena as causes in Collins 2018; Collins 2019, ch. 3.)

Organizations can also be part of causal explanations if we use a *process* theory of causal explanation. On these theories, the causes of some effect are the internal elements of the process that pre-empted the event—where that process might be a flow of energy, or might be the transmission of a modification in an otherwise consistent structure, or might be the conservation of a quantity (like mass-energy, linear momentum, or charge) within a given object over time (Schaffer 2016). For organizations to be potential causes on the process view, organizations would have to be concrete particulars: things through which energy can flow,

modifications can be transmitted, or mass-energy can be conserved. In Chapters Two and Three, I will develop and defend a view on which organizations are concrete particulars that have members as material parts. Lines can be drawn around them via their structures, as outlined in Section 1.2. They are thus amenable to analysis as causes, from within the process theory of causes. Naturally, both the counterfactual and process theories of causation are open to objections. But they demonstrate the breadth of causal-explanatory theories that accommodate organization-level phenomena as causes.

Functionalist Interpretivism

A final reason for realism about organizations is that they are agents. I'll assume that if we should be realists about anything, we should be realists about agents. I have argued elsewhere that collectives are agents—where 'collectives' are to be understood as entities composed of agents united under a group-level decision-making procedure that can process moral reasons (Collins 2019, ch. 6; see similarly List and Pettit 2011; Isaacs 2011; Hess 2018b; Hindriks 2018). As already explained, organizations are a type of collective agent: they are large, structured, labour-divided, governed, hierarchical collective agents. So, if my arguments elsewhere are correct, then it follows that organizations are agents.

It is worth briefly recapitulating the reasons for believing in organizations' agency, for two reasons. First, many working on collective agency (including my earlier self) do not sufficiently distinguish between *functionalist* and *interpretivist* reasons for believing in collectives' agency. This distinction is important, because philosophers of mind (and others) tend to fall on different sides of the functionalist-interpretivist divide. If organizations were agents in an interpretivist sense but not a functionalist sense, then any argument for collective agency that relied on interpretivism would alienate those who have independent reasons for believing in functionalism—and vice versa, if organizations were functionalist agents but not interpretivist ones. Thus, my aim here is to show that both accounts of agency render organizations agents.

The second reason for recapitulating reasons for believing in collectives' agency is that *organizations* specifically are particularly clear cases of collective agents. Thus, even if one doubts that collectives in general are agents (for example, if one was unconvinced by my or others' earlier arguments), then one may still be convincible that *organizations specifically* are agents. That's what I aim to do here.

What are the functionalist and interpretivist accounts of agency? I will start with functionalism. Within social ontology, functionalism about agency has been defended by Bryce

Huebner (2014), Gunnar Björnsson and Kendy Hess (2016), and David Strohmaier (2020). It's a popular account of mental states in general—and attitudes towards propositions in particular—throughout philosophy of mind. The characteristic feature of functionalism is that propositional attitudes are characterised by their functional profile. That is: attitudes such as belief, desire, hope, regret, and so on, when taken towards propositions, are characterised by the function that those attitudes play within some system. This system is the agent or entity that bears the attitude. The function of an attitude includes causing, grounding, or blocking other attitudes and behavioural dispositions. For example, we attribute to me the attitude of belief towards the proposition 'the ice cream is in the freezer,' I have an attitude that *causes* me to perform the behaviour of going to the freezer when I want ice cream. In this way, propositional attitudes are characterised by their relations to one another, and to behaviour.

By contrast, interpretivism has been defended within social ontology by Christian List and Philip Pettit (2011)[7] and Deborah Tollefsen (2015). Under this approach, the key question for agency is whether it's possible (or rational, or reasonable) for an external observer to adopt the 'intentional stance' towards an entity. The intentional stance is a mindset in which one views the entity's behaviours as explained by its beliefs, desires, hopes, fears, regrets, and so on. This stance is rational if it does a good job of explaining and predicting the observed behaviour of the entity. For example, if an observer sees me take the ice cream from the freezer, they might interpret me as desiring ice cream and believing it's in the freezer. After observing me for a few days, they might predict that I'll perform this behaviour in all future days after dinner. The (rationality of) ascriptions determine or even constitute my desiring ice cream after dinner and my believing it's in the fridge.

A key distinction between functionalism and interpretivism, then, is that the functionalist focuses on mechanisms internal to the agent, while the interpretivist focuses on behaviours that are observable from external to the agent. The functionalist also includes behaviours in their

[7] List and Pettit appeal to interpretivism at some points (e.g., 2011, 11, 13, 23) and to functionalism at other points (e.g., 2011, 171). However, at one point they say "the performance [behaviour] itself should dictate the representations and motivations we ascribe to the agent." (2011, 28) This places them more strongly in the interpretivist camp. I thank David Strohmaier for convincing me of this; see also Strohmaier 2020. Interpretivism differs from behaviourism: I take behaviourism to explain mentality just in terms of an entity's behaviours, while interpretivism explains mentality in terms of how an idealised observer would interpret those behaviours.

characterisation of an attitude's functional role, but the primary focus is on an attitude's relation to *other attitudes*, and in any case the behaviour is taken to be internal to the system, with no observer necessary. As I hope is clear, both functionalism and interpretivism have their merits.[8]

In the case of organizations, functionalism and interpretivism usually rise and fall together: an organization that's an agent by functionalist lights will almost always happen to be an agent by interpretivist lights, and vice versa. Thus, for it to be sensible for an observer to adopt the intentional stance towards an organization over time, it will usually be the case that the organization has certain internal mechanisms (such as mechanisms that ensure the organization behaves in ways that are roughly consistent across time, as emphasised by List and Pettit (2011)). Meanwhile, an organization's operation of rationality-producing internal mechanisms will almost always happen to result in outward organizational behaviour towards which it is appropriate for an observer to take the intentional stance. For purposes of figuring out whether real-world organizations are agents, then, we often don't need to pick between functionalism and interpretivism. Partly because of this, I will make things difficult for myself: I'll assume that an organization must meet *both* criteria in order to count as an agent.

There's another reason to insist that both functions and interpretations are important for organizations' agency: in the case of organizations, it's rather unclear where internal mechanisms end and observable behaviour begins (Strohmaier 2020). For example, are discussions between board members an internal mechanism, or an observable behaviour? If we endorse a combined 'functionalist-interpretivist' account (which demands an agent has *both* internal mechanisms *and* outward behaviour), then we don't need to draw a sharp line between internal mechanisms and observable behaviour.

What's more, we should require both outward behaviour and internal mechanisms because doing so gives us a stronger argument against the sceptic of organizational agency. The higher a bar we force organizational agents to meet, the less room there is for complaint

[8] Both also have problems. Arguably, each theory countenances too many agents. For functionalism, collective agents are sometimes taken to be counterexamples (Block 1978); another would be complex robots. For interpretivism, it seems odd that my mental states are determined by, or even constituted by, someone else's ascriptions. It seems there must be constraints on the interpreter's ascriptions, if the ascriptions are to determine my agency. I will assume such problems don't undermine understanding organizations as *agents*, though in Chapters Four and Five I will address nearby worries about organizations' *blameworthiness*, namely, the worry that organizations lack the phenomenology necessary for blameworthiness.

that we are being too permissive by including organizations in the category 'agent.' At the same time, requiring organizations to meet *both* the functionalist requirement *and* the interpretivist requirement enables us to retain a clean theory of organizational agency. The two requirements do not contradict one another, and nor are we treating them as disjuncts (which would result in a somewhat disunified theory, with some organizations being 'functionalist agents' and others being 'interpretivist agents'). Instead, the two requirements are layered neatly on top of each other. For all these reasons, I'll assume an organizational agents must have *both* internal mechanisms *and* rationalizable external behaviour.

On the functionalist picture, the features of organizations that render them agents are that their structure *coordinates* divided labour via *rules* and *hierarchical command relations*, guided by a *collective decision-making procedure*. While other kinds of collectives have the last of these features (a decision-making procedure), they do not necessarily coordinate divided labour via rules and hierarchical command relations.

Arguably, the decision-making procedure is enough to produce collective-level agency (assuming the procedure abides by rationality constraints). But the other three features (coordinating divided labour, rules, hierarchies) are not present in all collective agents, and they boost the case for the agency of organizations specifically. These features ensure that—by and large, and about as reliably as individual agents—organizations have large webs of mutually-consistent beliefs, preferences, and other propositional attitudes. Such large webs of mutual consistency are needed on the functionalist picture, because this picture characterises particular propositional attitudes via their causal (and other) relations to other propositional attitudes. Organizations' coordinated division of labour, rules, and hierarchies ensure that relations of consistency and mutual inter-definability hold between the organizations' various propositional attitudes. An organization's decision-making procedures, labour divisions, rules, and hierarchies are the mechanisms by which the web of mental states are produced.

What about interpretivism? Notice that organizations (perhaps unlike other collectives) are composed of people who *realise a structure*. This allows us to be clear about which behaviour belongs to the people qua private individuals, and which behaviour belongs to the organization—which ultimately allows us to attribute actions to the organization. The structure enables us to do this for two reasons. First, the structure provides boundaries to the organization: either a person is a role bearer or she is not; only the former's behaviour should be attributed to the organizational entity towards which we adopt the intentional stance. Second, the structure contains *roles*. It is only when a person is not just a role-bearer, but also a role-bear who is *performing her role*, that her behaviour should be attributed to the

organizational entity towards which we adopt the intentional stance. Thus, organizations' nature helps us to pick out the entity towards which we adopt the intentional stance.

Most importantly, when we adopt the intentional stance to an organization, we find behaviour that is understandable and predictable: it's understandable and predictable that banks will push legal limits for a profit and that states will permit environmentally dubious behaviour that boosts the short-term economy. Like under functionalism, the organization's decision-making procedures, labour divisions, rules, and hierarchies loom large here: these are the mechanisms that produce understandability and predictability in the organization's behaviour. Thanks to these mechanisms of organizations, we find that organizations behave in ways that make it rational for us to adopt the intentional stance towards them on a very regular basis. If interpretivism is true, this makes them agents. So we should view them as real.

1.4 The Questions

Having introduced organizations, their wrongdoing, and their reality, the stage is set for the book's substantive argument. As mentioned above, the book is divided into three parts: Part I is 'Metaphysics' (Chapters One to Three), Part II is 'Blameworthiness' (Chapters Four and Five), and Part III is 'Members' (Chapters Six and Seven). Parts I and II treat organizations as unified wholes—just as this chapter has done in providing a philosophical characterisation of organizations and three reasons for including organizations in our social ontology. Part III will consider the implications of organizations' wrongdoing for variously-placed members of organizations.

The rest of Part I is devoted to conceptualising organizations as material objects that are dependent upon members. I develop an account of the organization-member relation that doesn't *reduce* organizations to members, but that also acknowledges that members are crucial to organizations' performances. The aim is to produce an account that makes organizations *concrete material particulars*, so that we can more readily incorporate their causal powers into a naturalistic understanding of the world. We want a metaphysical account will enable us to hold organizations blameworthy in their own right, while producing the right kinds of moral implications for members.

Chapter Two will consider various ways of doing this that don't quite work. This will pave the way for Chapter Three's positive proposal, which walks the tightrope between the various pictures rejected in Chapter Two. On my proposed account, organizations are material objects that have members as material parts. Applied to the Banking Royal Commission, members of financial corporations include directors, non-manager employees (such as

financial advisors), and shareholders; but excludes intermediaries such as mortgage brokers (who offer products from more than one bank). Applied to the liberal democracies at issue in the Polluter Pays Principle, members include legislators, public servants, judges, and those with the right to vote in legislative elections; but excludes residents who lack the right to vote. On my picture, these individuals are the flesh-and-blood of organizations, even though organizations are not reducible to them.

Part II, 'Blameworthiness,' begins in Chapter Four by distinguishing recent moral-philosophical lenses on blameworthiness: the volitionist lens, the attributivist lens, and the aretaic lens. I explain how these lenses each apply to organizations. Most who have written on collectives' blameworthiness have presumed the volitionist lens is the only relevant lens. I will argue that social ontologists can draw more widely and liberally from moral philosophers' toolkits when it comes to our treatment of organizations. We can use the attributivist and aretaic lenses, as well as the volitionist lens. Drawing on Part I's metaphysical picture, I will argue that organizations should be attributed *both* the 'structural' wrongs that inhere in the organization as a whole *and* the 'individual' wrongs that are enacted by members-qua-members.

Chapter Five responds to an important objection to the idea that organizations can be wrongdoers, under any of the volitionist, attributivist, or aretaic lenses. The objection is that organizations lack the capacity for moral self-awareness, that is, awareness of their own wrongdoing. Unlike other philosophers who have defended collective blameworthiness, I believe the capacity for moral self-awareness is indeed presupposed by attributions of blameworthiness—and that this undermines a simple functionalist-interpretivist account of wrongdoers (even though I maintain that, as argued in this chapter, the functionalist-interpretivist account works for mere agents). I give a novel account of how organizations have the capacity for moral self-awareness. This account relies on the metaphysical story of Chapter Three. To foreshadow: the key innovation is that when members are performing their roles, certain pieces of their phenomenology *are* the organization's phenomenology. The organization 'inherits' phenomenology from its physical parts (members), just as a motorbike 'inherits' colour from its physical parts (engine, wheels, and so on). This allows organizations to have the phenomenology of moral self-awareness.

Part III, 'Members,' explains the implications of organizations' wrongs for members. Chapter Six catalogues three ways members can be implicated in any given organizational wrong: a member can be an 'enactor,' and 'endorser,' or an 'omitter.' I explain how divisions

of labour and hierarchical relationships can affect the type and degree of an individual's implication via each of these three mechanisms.

Chapter Seven provides a justification for members' reparative burdens when organizations are blameworthy for wrong. Importantly, Chapter Seven develops an account on which different members are liable to different shares of the reparative costs of their organizations' wrongs. Members who are implicated in the wrong (in one of the ways analysed in Chapter Six) are liable to a larger portion of the costs than non-implicated members. However, non-implicated members—including all present-day members of organizations that are blameworthy for historical injustice—can be liable to cost in virtue of having the capacity to repair relationships that were damaged by the organization's wrongdoing, or in virtue of their having benefitted from the organization's wrongdoing.