# Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions

Luciano Floridi

Oxford Internet Institute, University of Oxford, 1 St Giles, Oxford OX1 3JS, UK

The concept of distributed moral responsibility (DMR) has a long history. When it is understood as being entirely reducible to the sum of (some) human, individual and already morally loaded actions, then the allocation of DMR, and hence of praise and reward or blame and punishment, may be pragmatically difficult, but not conceptually problematic. However, in distributed environments, it is increasingly possible that a network of agents, some human, some artificial (e.g. a program) and some hybrid (e.g. a group of people working as a team thanks to a software platform), may cause distributed *moral* actions (DMAs). These are morally good or evil (i.e. morally loaded) actions caused by local interactions that are in themselves neither good nor evil (morally neutral). In this article, I analyse DMRs that are due to DMAs, and argue in favour of the allocation, by default and overridably, of full moral responsibility (faultless responsibility) to all the nodes/agents in the network causally relevant for bringing about the DMA in question, independently of intentionality. The mechanism proposed is inspired by, and adapts, three concepts: back propagation from network theory, strict liability from jurisprudence and common knowledge from epistemic logic.

This article is part of the themed issue 'The ethical impact of data science'.

# 1. Introduction

*Collective responsibility*[1] is a rather familiar concept, as old as the *Old Testament*.[2] According to it, a whole group of people is held responsible for some of its members' morally loaded (usually immoral) actions, sometimes even when the rest of the group has had no involvement at all (not even passively) in such actions. Equally well-known concepts are those of *shared responsibility*, *social* or *group actions* and (the theory of) *unintended consequences*. When these and similar phenomena are understood as being entirely reducible to the sum of (some) human, individual and already morally loaded actions, then the allocation of moral responsibility, and hence of praise and reward, or blame and punishment, may still be questionable and practically quite difficult, but it is not conceptually problematic. It lies squarely on the shoulders of all the individuals involved, totally, proportionally, in combination or perhaps not at all (exculpation). However, in distributed environments, it is increasingly common that a network of agents—some human, some artificial (e.g. a program) and some hybrid (e.g. a group of people working as a team thanks to a software platform)—may cause morally good or evil (henceforth loaded) actions through local interactions that are not, in themselves, morally loaded but neutral.[3] In a previous article,[4] I defined such a phenomenon as *distributed moral actions* (DMAs). I shall not elaborate on that analysis here because the focus of the following pages is rather a *consequence* of DMAs: what happens to the allocation of *responsibility* when we are dealing with DMAs arising from morally neutral interactions of (potentially hybrid) networks of agents? In other words, who is responsible (*distributed moral responsibility*, DMR) for DMAs? This is the question I wish to address in this article.[5] In §1, I shall clarify why ethics usually disregards DMR. I shall argue that it is mainly because ethics focuses on intentionality, which is, of course, not relevant in DMAs, the source of DMRs. Put simply, the reasoning is that, without intentionality, there is no DMA, and hence no DMR. This raises the question whether an ethics without intentionality may be meaningful at all. So, in §2, I introduce a simple sandbox[6] that will help clarify how good and evil may be brought about even without any reference to (or indeed presence of) agents' intentionality, and why an ethics without intentionality is not only possible but actually a necessary complement to the ethics of intentional actions. Without it, it may be virtually impossible to understand DMR. After this preparatory work, in §3, I introduce a mechanism to attribute DMR to a network of agents. The hypothesis is that a multi-agent system (from a whole society to just a group of agents, some of which may not be human, e.g. a group of bots interacting online) may be correctly interpreted as being equivalent to a multi-layered neural network. This is not very different from Plato's view in the *Republic*, where the individual and the city are discussed as the micro- and macro-level contexts at which actions take place. The network interpretation enables one to understand DMAs as the result of neutral interactions among the nodes of the network (*forward propagation*), and therefore allocate and indeed manage DMR in terms of *back propagation*

[1]The standard references are [1–4].

[2]See Jer 31:29 'The fathers have eaten a sour grape, and the children's teeth are set on edge'. And then Ezek 18:1–4 'The word of the Lord came to me: What do you people mean by quoting this proverb about the land of Israel: "The parents eat sour grapes, and the children's teeth are set on edge"? As surely as I live, declares the Sovereign Lord, you will no longer quote this proverb in Israel. For everyone belongs to me, the parent as well as the child—both alike belong to me. The one who sins is the one who will die.'

[3]The case in which the actions of the agents are morally good (morally loaded positively) but once aggregated cause evil effects (morally loaded negatively) is not discussed in this article, because the mechanism to locate responsibility in the neutral case can easily be extended to this 'loaded' case.

[4]See [5]. The two articles form a diptych, but they do not presuppose knowledge of each other.

[5]Note that [6] contains no discussion of responsibility, accountability or liability. Royakkers [7] discusses responsibility only 'forward', in terms of obligation/commitment to do something, i.e. in terms of a logic of 'seeing to it that' an action or a state is implemented (exercise of responsibility), not 'backward', in terms of blame/praise for something that has been done (attribution of responsibility). The latter is the topic of this article.

[6]I use the term here in analogy to its technical meaning in software development, where a sandbox is a safe testing environment that isolates untested code changes and experimentations from the production environment or repository. A sandbox replicates the minimal functionality needed to test the programs or other code under development.

to all the agents in the network that bring about the DMA. I shall then introduce and *adapt*[7] two more concepts needed to make sense of DMR: *strict liability* (borrowed from jurisprudence)[8] and *common knowledge* (borrowed from epistemic logic). Let me hasten to add that especially the first concept provides more an inspiration than a template for the proposed analysis.[9] Two illustrative examples will close that section. In §4, I shall comment on the mechanism introduced above by discussing two features of the analysis just developed, two objections to it, and two challenges facing it. In the conclusion, I shall highlight how the approach to DMR defended in this article shifts the focus from an ethics of responsibility based on individuals' intentional actions and oriented towards individual punishments and rewards, especially for legal and religious reasons (e.g. retributive justice or afterlife), to an ethics of responsibility based on groups' interactions and oriented towards environmental harm and welfare.

## 2. Why classic ethics does not focus on distributed moral responsibility

It is common to treat moral evaluations as *monotonic*,[10] in the following sense. If something is evil, it remains evil, even if it happens to lead to something morally good (henceforth simply good). This is a major reason why we argue that a good end does not justify evil means, and why accidental good consequences are not ground for praise. Likewise, if something is good, it remains good, even if it happens to lead to something evil. This is a major reason why we promptly excuse, and may even praise, people who cause some evil, if their intentions were genuinely good. Such a monotonic stability is shared by both deontological approaches, where it is admittedly more brittle (the *pereat mundus* approach), and consequentialist approaches, where it is actually more flexible (see the tension between act and rule utilitarianism, for example). Most importantly for our context, in ethics we often assume that what is morally neutral[11] remains neutral: if actions $a$ and $b$ are morally neutral, then their combination $C = a + b$ not only does not but cannot acquire a negative or positive moral value. Such a position is not incoherent, but it is criticizable in terms of a *modus tollens*. Morally loaded actions do occur as a result of morally neutral actions—this is the whole point of the tragedy of the commons,[12] for example—but the view that this is not the case should not be interpreted merely as a mistake, but more significantly as the correct consequence of a premise, which in itself is mistaken and should be replaced. The premise is that the ethical discourse should focus entirely and only on the intentional nature of actions. It is this exclusive focus on intentionality that makes it very difficult for standard ethics to deal with the attribution of DMR. Let me clarify.

[7]'Adapt' rather than 'adopt' because, as it will become clear, I refer to 'strict liability' only as a source, and not as an importable concept, for the formulation of strict or faultless responsibility.

[8]One of the anonymous reviewers rightly pointed out that 'All in all, I reckon that a more appropriate legal formula to convey the idea of the author borrowing the notion from jurisprudence, can be "faultless responsibility".' I agree, hence the title of this article. But I also noted that 'faultless responsibility' is a much less common concept than that of 'strict liability'. A quick search on Google, for example, returns 458 results for the former and about 3 840 000 results for the latter. Since in both cases it is only a matter of mere conceptual inspiration, in this paper, I took the liberty of keeping the original formulation. The reader is invited to switch to 'faultless responsibility' whenever this is deemed preferable.

[9]Strict liability is historically invoked and used in the legal field for two different reasons: (i) in order to assign liability for faultless behaviours as a result of an objective, either direct or indirect, process of causation and (ii) in order to assign liability as a result of a risk allocation despite a verified process of causation. Here I am mainly interested in the first meaning of strict liability (as part of the process of causation). I shall come back to the issue of risk allocation in the last part of the paper, but only tangentially.

[10]In logic, the *monotonicity* of entailment is a property of any logical system according to which the premises of a valid entailment may be freely extended with additional premises without making it invalid. In mathematics, a function or quantity is said to be *monotonic* if it varies in such a way that it either never decreases or never increases. The two senses are strictly related, since they point towards invariance under changed circumstances, but I am using 'monotonic' in the more mathematical sense of neither more nor less morally loaded than it was before the variation.

[11]By morally neutral, I mean here either not morally charged at all or below a threshold of moral relevance (virtually amoral). This specification is crucial since one may argue that almost any action shows at least traces of moral value. Moreover, given the right circumstances, any action may become morally loaded. Alice scratching her left foot may cause unspeakable evil if one can imagine the right chain of causes.

[12]See [8,9]. On the digital version of the tragedy, see [10]. Other examples of distributed moral actions are presented in [5].

Intentionality is not closed under causal implication, whether *direct* or *distributed*. In the *direct* case of non-closure, it is not the case that, if Alice means to cause $a$, and $a$ causes $b$, it follows that Alice means to cause $b$. In the *distributed* case of non-closure, it is not the case that, if Alice means to cause $a$, and Bob means to cause $b$, and $a$ and $b$ cause $C$, it follows that Alice and Bob mean to cause $C$. To be more precise,

(i) $\neg$ [[[A means to cause $a$] $\wedge$ [$a$ causes $b$]] $\rightarrow$ [A means to cause $b$]]
(ii) $\neg$ [[[A means to cause $a$] $\wedge$ [B means to cause $b$] $\wedge$ [$a \wedge b$ cause $C$]] $\rightarrow$ [AB means to cause $C$]]

Both (i) and (ii) are correct. But precisely because intentionality is not closed under direct or distributed causal implication, the assumption of the intentionality of an action as a necessary condition for an ethical evaluation of it leads to the oversight of DMAs and responsibilities. The reasoning may be summarized in the following steps:

(1) the classic emphasis is on the allocation of individual punishments and rewards, especially for socio-legal and religious reasons (e.g. retributive justice or afterlife), not on the allocation of risks of environmental harm and opportunities of environmental welfare (more on this later; see also [11]);
(2) so the allocation in (1) must focus on the attribution of individual *responsibility*;
(3) (1) and (2) lead to the identification of individual *intentionality*; it would be counterproductive to attribute responsibility, and hence allocate blame or praise, punishments or rewards, if the agents' actions were not intentional, because such attribution would then be arbitrary and indistinguishable from a mere random allocation, which would defy the purpose of blame or praise, punishments or rewards, insofar as these are meant to modify and guide possible choices and actions for the benefit of the individuals involved and their society;
(4) but we have seen that intentionality is not closed under causal implication: when a DMA $C$ is in question, such resulting action is not intentional;
(5) but then it follows that no agents (say, neither Alice nor Bob), whose neutral actions bring about $C$, are treatable as being responsible for $C$;
(6) therefore neither Alice nor Bob can be fairly punished or rewarded for $C$; and
(7) yet evaluating individual agents and their moral lives is the whole point of an ethical analysis; therefore, DMR is a phenomenon on which standard ethics does not focus.

A direct consequence of (1)–(7) is that standard ethics either ignores DMAs and responsibilities or seeks to reduce both to non-distributed versions of individual morality of intentional actions. Both strategies are unsatisfactory. Ethics is not only a matter of evaluating agents and their intentional actions, but also a matter of evaluating the states of the receiver of the action (the patient affected), and hence of influencing the relevant groups of agents whose aggregated actions lead to such states. If what drives the analysis is the question whether the patient affected is morally better or worse off after an action has taken place, then intentionality may still be very relevant, but it is no longer a necessary condition, and it becomes crucial to understand how one may allocate DMR for DMAs that emerge from entirely neutral actions, so that the right actions are facilitated, promoted, amplified and rewarded, and the wrong actions hindered, prevented, mitigated or punished in reparation.

## 3. The three approaches to ethics: agent-, action- and patient-oriented

To understand how an ethics without intentionality—what I have defined elsewhere as *mindless morality* [12]—may be possible, let me introduce now the sandbox I anticipated above. This is a simple environment that can help us to clarify how ethics may not be about the states of agents

| | | system states | | | |
|---|---|---|---|---|---|
| | | $S_1$ | $S_2$ | $S_3$ | $S_4$ |
| actions | $A_a$ | $S_2$ | $S_3$ | $S_2$ | $S_4$ |
| | $A_b$ | $S_1$ | $S_1$ | $S_2$ | $S_4$ |
| | $A_c$ | $S_1$ | $S_1$ | $S_4$ | $S_1$ |

**Figure 1.** Example of an elementary finite-state automaton.

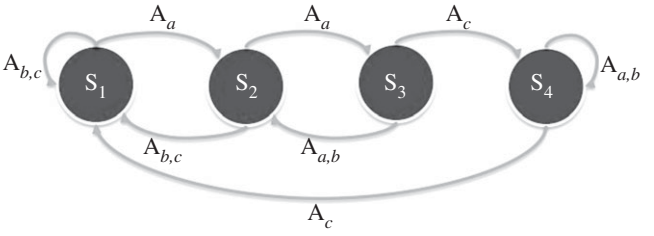


**Figure 2.** Graphical representation of an elementary finite state automaton.

or their intentional actions, but about the states of the environments affected by agents and any of their actions.

Consider a finite state automaton (FSA [13]). An FSA is the sort of logical scheme that describes how a vending machine works. Think of it as a system that consumes actions as inputs to deliver changes of states as outputs, e.g. a payment and a choice of drink in order to deliver the chosen drink. An FSA, as a simple scheme of action, is defined by

1. a finite set of **states,** for example four $S$: $\{S_1, S_2, S_3, S_4\}$;
2. a finite **alphabet** (set) of input/actions, for example three $A$: $\{A_a, A_b, A_c\}$;
3. a transition function $f$: $S \times A \rightarrow S$;
4. a **start state** $S_1 \in S$; and
5. a set of **acceptable states** $F \subseteq S$.

Figure 1 illustrates the simple example of an FSA just introduced. It is read by checking which action-input (e.g. $A_b$), given a system's state (e.g. $S_2$), puts the system in which output-state (in this case $S_1$). Figure 2 illustrates figure 1 graphically.

Because an FSA is a basic example of a system that moves from one state to another as a result of action-inputs, it is the bare minimum sufficient to clarify some elementary features and dynamics of a very simple system, a sandbox that makes it possible to identify some crucial issues in the analysis to be developed. Remember that this sandbox is not a model (it does not represent an existing system), a blueprint (it does not indicate how to build a system not yet existing) or a thought experiment (it is not a mere matter of imagining a non-contradictory system, because it comes with very concrete formal constraints) for an ethical system. It is a simplified environment to test some ideas, in the following way.

In virtue ethics, we focus on the nature (e.g. their characters, intentions, inclinations, choices or plans) of the agents that implement the action-inputs that lead to the transitions in the system. In deontological and consequentialist approaches, we may disregard the nature of the agents to focus on the nature of the action-inputs leading to the transitions in the system, thus moving from an agent-oriented to an action-oriented approach. In environmental contexts, we may disregard both agents and their actions to focus on the features of the system that we want to see pursued or avoided. We, thus, move to a patient-oriented approach. It is clear that any ethical analysis,

from the very elementary one I just introduced to the most complex and realistic ones, presents three points of 'pressure' where a difference can be made to good and evil. To promote good and eradicate evil, one may seek to change the nature of the agents, of their actions or of the states of the patients (these are inclusive disjunctions). With an analogy, the ethical discourse may focus on the cook, on the cooking or on the cooked. In [12], I have argued that a patient-oriented approach is not only defensible but in some cases preferable for the development of our ethical discourse. I shall not rehearse the reasons provided there in support of such a position, because what matters here is that, once this tripartite distinction is available, it becomes clear that an intention-based (agent- or action-oriented) ethics is not the only one available, and indeed that an ethics of state transitions, independent of the intentions of the agents involved, can provide a full account of DMR. All we need to assume is that, according to an axiological analysis, some states of the system are morally better than others and hence worth pursuing for the sake of the system itself. Understanding how they are brought about (the nature of the actions), and according to which plans or intentions (the nature of the sources of the actions), may be crucial to answer significant moral questions (including the classic 'who should I be?'), but it is not strictly necessary (it is not a *sine qua non*) to evaluate whether the *moral patient*, the receiver of such actions, is morally better or worse off. One may imagine a scenario (a level of abstraction, see [14,15]) in which no information[13] is available about the agents involved or their actions. If the only perceivable changes concern the state transitions of the system affected, as described in the sandbox, one would still be able to provide an ethical assessment. Note, however, that the point is not to develop such an axiology here.[14] In this context, we can just assume that one is possible and indeed available. In our sandbox, for example, we may simply stipulate that our axiological analysis determines that $S_1$ is a morally negative state (evil), that $S_2$ and $S_3$ are neutral and that $S_4$ is a morally positive state (good).

It seems clear that it is perfectly fine to talk about moral states independently of agents' intentionality and the specific moral nature of their actions. And this means that we can finally ask the question motivating this article: if a DMA fails to bring about an increase in the moral value of the system—in our sandbox: if a DMA fails to move the system to $S_4$ and keep it there, who is responsible for it? Answering this question is the task of the next section.

## 4. How to allocate distributed moral responsibility

Attributing moral responsibility, irrespective of intentionality and information about the nature of the agents involved and their actions, means focusing on which agents are causally accountable for (i.e. contributed genetically to bring about) a morally distributed action *C*, rather than whether agents are fairly commendable or punishable for *C*. This means talking about 'responsibility' in the aetiological sense of being the source of (causally accountable for) a state of the system, and therefore, as a consequence, of being morally answerable (blameable/praisable) for its state. This may lead to, but it is independent of, legal liability, in the sense of being subjectable to sanction or reward.[15] And it either grounds or is independent of the concept of responsibility understood as being in charge of something and hence seeing to it that something happens or does not happen. The only assumption required is that the agents causally accountable can learn from, and modify, their behaviour. In other words, we only need to assume that the agents in question are autonomous (in the minimal sense that they are in charge and regulate their own actions, at

[13]In this article, I presuppose an informational analysis of knowledge (see [16,17]), yet nothing depends on this.

[14]I have developed an axiological analysis that is e-nvironmentally (hyphen intended) oriented in [12].

[15]It is clear that, at this point, a more careful analysis of causal relations is needed, yet this is left to a future work to see which philosophical theory of causality better fits the analysis provided in this article. Here, the interested reader may note that, on the legal side, the starting point is the classic [18], and on the philosophical side I would recommend [19,20]. I have inclined towards an analysis of causality as a purely informational interpretation of continuous events at a given level of abstraction (LoA) chosen for a purpose in [21], where I support a concept of 'sufficientization' ($x$ is a sufficient condition for $y$) at a given LoA for a purpose. This is close to the approach taken by Hart & Honoré [18] in terms of 'purpose of the inquiry' and 'sufficient intervention'; see the Introduction, §1, entitled 'Causation' in [22]. On the analysis of the 'failure of causation' in jurisprudence see [23, pp. 68–72].

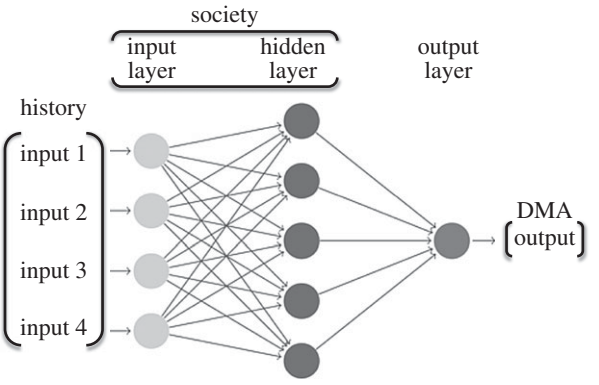


**Figure 3.** A multi-agent system as a multi-layered neural network.

least to some significant extent), can interact with each other and their environments and can learn from their interactions (can change the rules according to which they behave, again, at least to some significant extent). If they satisfy these three necessary and sufficient conditions (and most humans as well as some artificial or hybrid agents can; for a detailed analysis, see [12,24]), then they give rise to a *multi-layered neural network* that can learn its appropriate internal representations and hence any arbitrary mapping of input (the preceding history and context) to output (DMA), and improve its behaviour. Figure 3 provides an elementary illustration of the various elements in such a network. In it, the input of the network, on the left, is labelled 'history'. It represents past information (e.g. circumstances, choices already made, past plans and any information already available). The network in the middle—the two layers of nodes labelled 'society'—transforms such input into an output. The output, on the right, is the DMA. Forward propagation (from left to right) is how a DMA is outputted, while back propagation (right to left) is how DMR can be attributed, in view of an improvement of the state of the system affected by the outputted DMA.

In forward propagation, the agents in the network output, as a whole, a distributed action that is morally loaded, by activating themselves and by interacting with other agents according to some specific inputs and thresholds, in ways that are assumed to be morally neutral. In such a distributed context, it no longer matters which agent does what or why. All that matters is that the change in the system caused by the DMA is good or evil and, if it is evil, that one can seek to rectify or reduce it by treating the whole network as accountable for it, and hence back propagate responsibility to all its nodes/agents to improve the outcome. The cycle ends when the output is satisfactory, according to the chosen axiological analysis. Real neural networks achieve this stability by de-/activating specific nodes (agents) in the network and/or by finding the derivative of error with respect to each weight of their links and then subtracting this value from the weight value, until the desired outcome is obtained. In a social network, this is achieved through hard and soft legislation, rules and codes of conducts, nudging, incentives and disincentives; in other words, through social pushes and pulls. It follows that the analysis of DMAs requires the following steps:

(a) identification of the DMA $C_n$;
(b) identification of the network $N$ causally accountable for $C_n$ (forward propagation);
(c) back propagation of moral responsibility to make each agent in $N$ *prima facie* equally and maximally responsible for $C_n$;
(d) correction of $C_n$ into $C_{n+1}$; and
(e) repetition of (a)–(d) until $C_{n+1}$ is axiologically satisfactory.

Steps (a) and (b) are aetiological. They are conceptually uncontroversial, although their implementation may be challenging in practice, and perhaps sometimes just impossible. Step (c)

may surprise the moral philosopher but not the legal scholar, because it *resembles strict liability*.[16] Especially, but not only,[17] in tort law, strict liability is the legal responsibility of one or more agents for the damage or loss caused by their acts or omissions, regardless of their culpability, where the latter is defined in terms of intentionality of the action, possibility to control it and lack of excuse. Under *strict* liability, there is no requirement to prove fault, negligence or intention.[18] Interestingly, strict liability is most commonly associated with damage caused by animals and defectively manufactured products. This is not accidental. The reference to animals is linked to the need for a mindless morality, in a context where keepers of the animals become strictly liable for their animal agents, to whom it is difficult to attribute moral intentions in the ordinary sense of the expression. And the design perspective is consistent with the patient-oriented approach adopted at the beginning of this article, which looks at the receiver of the action as a system that is being designed by the agents issuing the actions. If the design is poor and the outcome faulty, then all the agents involved are deemed responsible. One needs to show that some evil has occurred in the system, and that the actions in question caused such evil, but it is not necessary to show exactly whether the agents/sources of such actions were careless, or whether they did not intend to cause them. It is important to note that strict liability has given rise to corporate liability in criminal law. This establishes how far a corporation, as a legal person, can be liable for the acts and omissions of the natural persons it employs. Yet, note that this is not how I intend to use the 'family resemblance' between 'strict moral responsibility' and strict liability, because I intend to keep the same scope of applicability (all individual agents involved), not shift it (the network). Here, faultless responsibility remains 'theirs' (agents') not 'its' (network's). All that is needed in (c) is the mechanism of 'responsible by default' or *poena sine culpa*, to invert the Latin phrase [28].

Step (d) may require an overridability clause. Some nodes may share different degrees of responsibility, including none at all, if an agent is able to show no involvement in the interactions leading to *C*. I shall return to this point below, when discussing two examples.

Finally, step (e) may not be required, if the presence of back propagation of DMR is known to all the agents involved, and this knowledge prevents an evil DMR from being outputted in the first place. If all the agents know that they will all be responsible for *C*, it is more likely that *C* may not occur, as they may restrain themselves and each other. This is social pressure, and I shall say more about it in the conclusion. To achieve such a preventive result, a simple mechanism of so-called *common knowledge* may be sufficient. *Common knowledge* of $p$ occurs in a group of agents $G$ when all the agents in $G$ know that $p$, they all know that they know that $p$, they all know that they all know that they know that $p$, and so on *ad infinitum*. This is achievable through a *public announcement*, an informative event that updates all the agents in $G$ about $p$ in a way perceivable by all agents. *Common knowledge* through *public announcement*—two concepts in epistemic logic well known to legal scholars in terms of common and public knowledge of the law—could be pursued to put all the agents in $G$ in charge of *C*, and thus increases the chances that they may be able to prevent or modify an evil *C* or at least not participate at all in its delivery. This is the substantive aspect in which DMR is very different from collective responsibility.[19]

[16] The history of strict liability is long and complicated, not least because it seamlessly interacts with the history of moral responsibility in mutual interchanges of conceptual modifications. Two texts helpful in mapping the development of the concepts are Vandall [25], who supports the extension of strict liability far beyond the area of products liability, somewhat in line with the ethical argument developed in this article; and Epstein [26], who supports a purely aetiological analysis for the evaluation of strict liability, another point on which I agree in this article.

[17] For an evaluation of the controversial extension of strict liability to criminal law, see [27].

[18] This is not trivial, since sometimes being innocent does not mean not being liable. There are laws that stipulate liability regardless of any fault, that is, even if the person is able to prove that he or she is innocent, that person will be held responsible for system failures. A typical example is the operator of a nuclear plant, which will be held responsible for any damage caused by the nuclear plant. If people engage in dangerous (although lawful) activities that could harm the population, they are held responsible for any damage that occurs as a result of such activities. In cases of strict liability, the defendant is allowed to prove that he or she is innocent, which then leads to an exemption of liability.

[19] Note that this use of common knowledge and public announcement should be clearly distinguished from the issue of 'knowledge of the law'. As one of the anonymous reviewers rightly remarked: 'a "public announcement" of the law does not necessarily amount to a "public knowledge" of the law: a legal public announcement refers to the public access to the sources of law (and not necessarily to its knowledge, which normally requires the interpretation of the law)'. Here I am referring to the two concepts of 'common knowledge' and 'pubic announcement' as they are used in epistemic logic.

Two examples may now help illustrate the previous analysis. They both come from The Netherlands, and are known as 'the three cyclists' and 'the four boats'.[20]

The Netherlands is famous for its friendly approach to bicycles and cyclists. The Dutch road traffic regulations allow at most two cyclists to ride next to each other, if they do not endanger others. What happens if a third cyclist joins them? Each binary action, describable as 'Alice and Bob cycling together', is considered to be safe, that is, morally neutral in the vocabulary of this article. But if several binary actions take place, hazard may emerge. The action $C$ that comprises more than two people cycling together is morally loaded negatively. In a context where there are no DMAs and DMRs, one may assume that only the third person who joins the other two already cycling together is to be held responsible, that is, only the one most left (the Dutch drive on the right side of the road). Yet this is not the case. In 1948, the Dutch Supreme Court ruled that each of them is to be held entirely responsible, because it is very easy for each of them to rectify the situation (HR 9 March 1948, NJ 1948, 370). This back propagation of responsibility means that all cyclists pay attention not to be cycling together in more than two, not just in terms of not joining a couple but also in terms of not being joined by a third cyclist.

Contrast this with a comparable yet different case concerning an equally famous aspect of life in The Netherlands, namely boats, rivers and canals. Dutch law allows up to three boats to be moored next to one another breadthways on the river (Merwede) outside harbours. In 1931, when a fourth ship was moored next to three others, the Dutch Supreme Court ruled that only the fourth ship was responsible, because it was much more difficult for the other three to rectify the situation than for the fourth that joined them (HR 19 January 1931, NJ 1931, 1455). In this case, a DMA led to a DMR, but the back propagation identified only one agent as responsible, even if the DMA required all four of them to occur.

The two examples remind us that understanding and insight will need to be exercised when back propagating strict forms of DMR, but they also show that this is both possible and ordinarily done.

## 5. Features, objections and challenges

Let me now comment on the previous proposal by highlighting two features, two objections and two challenges characterizing the mechanism outlined in the previous section. The first feature is its *uncommitted stance*. Interpreting DMA as being outputted by a network of agents—which could be as small as three people cycling together and as large as an entire society—enables one to design strict, back propagated, overridable DMR in ways that by-pass the classic intentionality hurdle. However, it says nothing about the axiology implemented. Recall that, in our sandbox, we merely stipulated the moral values of the four possible states of the system. This lack of commitment is a positive feature. The mechanism of attribution of DMR is neutral with respect to the actual moral evaluation of the output. It can work even to 'invert' a good outcome. In our sandbox, it would be the same mechanism that would make the system move from $S_4$ to $S_1$. This means—and this is the second feature—that an axiological analysis is unavoidable, because the design of the mechanism of attribution of DMR is actually part of the design of a society's *infraethics*, rather than of its *ethics*. An infraethics is the ethical infrastructure that, although not morally good or evil in itself, can facilitate or hinder actions that lead to good or evil states of the system. Which states should be implemented or not is up to an axiological theory to decide, but how easily they can be implemented is part of the infraethics, and the mechanism of DMR attribution plays a significant role in the latter.

[20]For an elegant formalization of both examples in terms of deontic modalities and logic of action, see [7]. For a hybrid example, not developed in this article, consider bots. In 2014, software agents known as bots completed about 15% of all edits on Wikipedia [29]. Such bots are approved by Wikipedia and they are, on the one hand, collaborative agents that depend on, and interact with, human users, who program them and can guide them (as well as switch them on/off), but, on the other hand, they are autonomous agents, which can work interactively and learn from their environment. They can take and execute decisions with or without human intervention and perceive and adapt to the context within which they operate. The allocation of responsibility when such bots make mistakes is clearly a matter of distributed actions. It will be the topic of another article.

Strictly responsibilizing the agents in the network that brings about a morally loaded change in the state of a given system may seem unfair and against their fundamental rights,[21] if no intentionality is involved. This is the first objection. And it is reasonable. The answer to it is twofold. On the one hand, some evil in the world, and the back propagated allocation of its faultless responsibility, is tragic, that is, it is indeed unfair: one is found (or, more often, finds oneself) responsible for *C* even if one (knows that one) could not have done anything to avoid or prevent *C*. It is what Tony Honoré analyses in terms of 'outcome responsibility', which holds even in cases of bad luck because, he argues, an attribution or assumption of responsibility is acceptable simply on the basis of an agent's intervention in the world [22]. On the other hand, and this is no longer biting the bullet, when the circumstances are not tragic, the lack of reference to intentionality is (at least partially[22]) counterbalanced by the presence of common knowledge, reached through public announcement, about the mechanism in place: agents are (or need to be) informed that a back propagation of strict DMR is implemented, in the same way as cyclists in The Netherlands are (need to be) informed that all three will be sanctioned if a third cyclist joins two already cycling together. In this way, the attribution of strict DMR is meant to play a significant role in preventing evil and fostering good, not in blaming or punishing agents for their morally unsuccessful actions.

The second objection questions whether the mechanism is realistic. Here, the response is that we already apply a blunt version of back propagation of strict DMR. This happens when we blame leaders (CEOs, Directors, Presidents, Prime Ministers, Generals and bosses of all kind) for the mistakes made by those whom they lead, even if they lack any information or intentionality about the latter's intentions or actions. Indeed, we ask them to play such a role also, if not mainly, because we lack a better way of allocating DMR. The mechanism suggested in this article is only a refinement of that approach, and this is why it is more, not less, realistic. Instead of blaming only some principal nodes/agents in the network, and often only one on the basis of some conventions—the vulgate states that with great power comes great responsibility—the suggestion is to allocate responsibility less coarsely, across all the relevant nodes/agents in the network. The advantage is that the more people who are going to be deemed responsible for some evil, the more likely it is that some of them will call for more caution to be exercised. Likewise, the more people share in the praise and rewards, the more likely it is that good will be pursued. This is the other side of the bonus culture.

The last point leads to one of the challenges I anticipated. The back propagation mechanism may promote risk aversion, and this may be a difficulty. If all agents in the network are made equally and fully responsible for the outputted morally distributed action—recall the three cyclists example—then it is possible that some of them, more prudent, will adapt more readily than some others, less prudent, to deal with such DMR. In the three cyclists example, some cyclists may decide never to ride in a couple, just to be on the safe side; and a third, imprudent cyclist joining a couple may force a more prudent member of the couple to leave. This is the same phenomenon that makes it possible for reckless drivers on a motorway to be safe, by relying on the restraint and extra care exercised by the majority of more careful drivers. The outcome is a promotion of a few cases of irresponsibility counterbalanced by many cases of extra cautiousness. This first challenge is due to the fact that the resilience of the network and its ability to improve its performance is not based on all agents in the network sharing an equal degree of risk aversion. As long as the network improves its output, the mechanism of back propagation of faultless responsibility makes it irrelevant whether some agents contribute more or less to the performance. This is in itself a problem of fair distribution of pressure: some agents will feel more in need to improve the overall outcome than others. However, through time, interactions between more

[21] On the distinction between relative and absolute human rights, see [30].

[22] It is fair to object that this appeal to common knowledge does not address different agents' relative costs of defection from the network. To take a corporate example, a highly successful executive can afford to challenge a boss's immoral policy proposal, or even quit the job in protest, than can a young employee. This is another reason why an intelligent and informed evaluation of the circumstances remains unavoidable. In terms of a sense of responsibility, individuals will feel differently about it precisely because of a difference in the relative costs.

prudent risk-averse agents and more imprudent risk-seeking ones should lead to an equilibrium that can be rectified, if the agents involved find it unsatisfactory, in any combination of three ways. First, the equilibrium could be improved by allocating further individual responsibilities, which have not disappeared. The agent/node that misbehaves may still be held responsible for any excessive risk-taking behaviour, i.e. behaviour that, in itself, is already morally loaded, as with the reckless driving example mentioned above. Second, incentives and disincentives may be designed to limit the effect of the lack of balance in risk-taking among the agents in the network. This would be the equivalent of introducing more finely measured cases of responsibility between the 'four boats' example at one end (the previous three boats have no responsibility) and the 'three cyclists' example at the other end (all cyclists have full responsibility) of the spectrum. This can be achieved by identifying circumstances in which DMR is back propagated proportionally to the ability of the agents to avoid the negative outcome. Think of a case in which one of the three boats could easily move, or neither of the two cyclists could possibly leave the group because a chain connects them and they are driving to a locksmith to break it when joined by a third cyclist. Third, social pressure from more prudent agents/nodes may constrain the more risk-prone behaviour of the less prudent agents/nodes. In our example, the two cyclists may firmly complain to the third one for having caused them to pay a fine.

The same social pressure may lead to the last challenge I wish to highlight. The back propagation of faultless responsibility may stifle innovation and support a culture that is too risk averse. If any agent in the network is fully responsible, morally, for what the network outputs as a DMA, then this may encourage some or perhaps even all agents to refrain from acting or even abandoning the network, if they can. In terms of the tragedy of the commons, nobody would use the commons, just in case using it even once led to full responsibility for its depletion. This is not a welcome outcome, for it would mean that opportunities would be missed, and resources would be wasted. In this case too, the design of proper incentives to encourage agents to take some reasonable and limited risks may be pursued. In economics, this could be a matter, for example, of insurance policies to hedge against liability. In an ethical context, such moral hedging may be provided by a better understanding of an agent's duties towards proactive care of the system affected.

## 6. Conclusion

In a world where the complexity and long-term impact of human–machine and networked interactions are growing exponentially, we need to upgrade our ethical theory to take into account the highly distributed scenarios that are becoming so increasingly common. Too often 'distributed' turns into 'diffused': everybody's problem becomes nobody's responsibility. This is morally unacceptable and pragmatically too risky. It is why, in this article, I have sought to provide a mechanism for the allocation of DMR caused by DMAs. Shared responsibility in international relations may have to become faultless responsibility.[23] In the course of the previous pages, I have explicitly adopted a design perspective. I have argued that a successful strategy to tackle the problem of DMR is to formulate a mechanism that, by default, back propagates all the responsibility for the good or evil caused by a whole causally relevant network to each agent in it, independently of the degrees of intentionality, informed-ness and risk aversion of such agents (faultless responsibility). The shift in perspective is from an agent-oriented ethics, which cares about the individual development, social welfare and ultimate salvation, to a patient-oriented ethics, which cares about the affected system's well-being and ultimate flourishing. To put it bluntly, this means shifting the focus from an agent's interest to a patient's harm. Our world may not need an ethics for Paradise and individual sins, but it definitely needs an ethics for Eden and environmental risks.

Competing interests. I declare I have no competing interests.
Funding. I received no funding for this study.

[23]For an analysis of shared responsibility in international law, see [31–33].

## References


1. French PA, Wettstein HK. 2006 *Shared intentions and collective responsibility*. Boston, MA: Blackwell Publishing.
2. French PA. 1998 *Individual and collective responsibility*, 2nd edn. Rochester, VT: Schenkman Books.
3. May L. 1987 *The morality of groups: collective responsibility, group-based harm, and corporate rights*. Notre Dame, IN: University of Notre Dame Press.
4. May L, Hoffman S. 1991 *Collective responsibility: five decades of debate in theoretical and applied ethics*. Savage, MD: Rowman & Littlefield.
5. Floridi L. 2013 Distributed morality in an information society. *Sci. Eng. Ethics* **19**, 727–743. (doi:10.1007/s11948-012-9413-4)
6. Olson M. 1965 *The logic of collective action; public goods and the theory of groups*. Cambridge, MA: Harvard University Press.
7. Royakkers LMM. 1998 *Extending deontic logic for the formalisation of legal rules*. Dordrecht, The Netherlands: Kluwer Academic.
8. Hardin G. 1968 The tragedy of the commons. *Science* **162**, 1243–1248. (doi:10.1126/science.162.3859.1243)
9. Hardin G. 1998 Extensions of 'The tragedy of the commons'. *Science* **280**, 682–683. (doi:10.1126/science.280.5364.682)
10. Greco GM, Floridi L. 2004 The tragedy of the digital commons. *Ethics Inform. Technol.* **6**, 73–81. (doi:10.1007/s10676-004-2895-2)
11. Floridi L (ed.). 2008 *Philosophy of computing and information: 5 questions*. Copenhagen, Denmark: Automatic Press/VIP.
12. Floridi L. 2013 *The ethics of information*. Oxford, UK: Oxford University Press.
13. Sipser M. 2013 *Introduction to the theory of computation*, 3rd edn. Andover, MA: Cengage Learning.
14. Floridi L. 2008 The method of levels of abstraction. *Minds Mach.* **18**, 303–329. (doi:10.1007/s11023-008-9113-7)
15. Floridi L. 2010 Levels of abstraction and the Turing test. *Kybernetes* **39**, 423–440. (doi:10.1108/03684921011036150)
16. Floridi L. 2004 On the logical unsolvability of the Gettier problem. *Synthese* **142**, 61–79. (doi:10.1023/B:SYNT.0000047709.27594.c4)
17. Floridi L. 2011 *The philosophy of information*. Oxford, UK: Oxford University Press.
18. Hart HLA, Honoré T. 1985 *Causation in the law*, 2nd edn. Oxford, UK: Oxford University Press.
19. Illari PM, Russo F, Williamson J. 2011 *Causality in the sciences*. Oxford, UK: Oxford University Press.
20. Illari PM, Federica R. 2014 *Causality: philosophical theory meets scientific practice*. Oxford, UK: Oxford University Press.
21. Floridi L. In press. Design as a conceptual logic of information. *Minds Machines.*
22. Honoré T. 1999 *Responsibility and fault*. Oxford, UK: Hart.

23. Pagallo U. 2013 *The laws of robots: crimes, contracts, and torts*. New York, NY: Springer.
24. Floridi L, Sanders JW. 2004 On the morality of artificial agents. *Minds Mach.* **14**, 349–379. (doi:10.1023/B:MIND.0000035461.63578.9d)
25. Vandall FJ. 1989 *Strict liability: legal and economic analysis*. Westport, CO: Quorum.
26. Epstein RA. 1980 *Theory of strict liability: toward a reformulation of tort law*. San Francisco, CA: CATO Institute.
27. Simester AP. 2005 *Appraising strict liability*. Oxford, UK: Oxford University Press.
28. Tully M. 2000 *Poena sine culpa?: Strict-liability-Sanktionen und europäisches Gemeinschaftsrecht*. Frankfurt am Main, Germany: P. Lang.
29. Steiner T. 2014 Bots vs. Wikipedians, anons vs. logged-ins. In *Proc. of the 23rd Int. Conf. on World Wide Web, Seoul, Republic of Korea, 7–11 April 2014*, pp. 547–548. New York, NY: ACM.
30. Pagallo U. 2013 Online security and the protection of civil rights: a legal overview. *Philos. Technol.* **26**, 381–395. (doi:10.1007/s13347-013-0119-6)
31. Nollkaemper A, Plakokefalos I. 2014 *Principles of shared responsibility in international law*. Cambridge, UK: Cambridge University Press.
32. Nollkaemper A, Jacobs D, Schechinger JNM. 2015 *Distribution of responsibilities in international law*. Cambridge, UK: Cambridge University Press.
33. Nollkaemper A, Jacobs D. 2013 Shared responsibility in international law: a conceptual framework. *Mich. J. Int. Law* **34**, 359–361.