HUMAN-MACHINE COMMUNICATION

Human-Machine Communication
Volume 10, 2025
https://doi.org/10.30658/hmc.10.4

Collectives in Cases of Human-Machine Interaction: A Pragmatic Defense of Human-Machine Collectives

Niël Henk Conradie[1], Hendrik Kempt[1], Nils Freyer[2], and Saskia K. Nagel[1]

1 Department of Technology, Society, and Human Factors, RWTH Aachen University, Germany
2 University of Applied Sciences, FH Aachen, Germany

Abstract

Humanity's future will include greater integration with different kinds of increasingly autonomous and interactive technological systems. Doing so is likely to confront us with reasons to adapt some of our currently accepted concepts and practices. We contend that among these is included a persuasive pragmatic reason to undertake just such an adaptation in our concept of collective responsibility and the practices attendant to it. The core of this pragmatic reason lies in how it allows us to respond to the issue of possible techno-responsibility gaps, which have been at the center of much of the attention and debate surrounding autonomous technologies. Our guiding thought is that we ought to consider a spectrum of interactiveness as an additional parameter when confronting such possible gaps. We contend that as one proceeds along the spectrum of increasingly depthful interactivity between human and machine, individual responsibility fails to provide the tools necessary for fully accounting for the role of responsibility, and we turn to the idea of collectivizing responsibility in such cases. We will claim that collectives can be instantiated by some agent-non-agent relationships, with the central condition being that this relationship reaches a meaningful level of interactivity. We take a requirement for this

Acknowledgment: This work was supported by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung (BMBF), now under the new name Bundesministerium für Forschung, Technologie und Raumfahrt (BMFTR)) as part of the projects BridgingAI (16DHBKI023), NeuroSys (03ZU1106EA), and VEREINT (16SV9108).

CONTACT Niël Henk Conradie • niel.conradie@humtec.rwth-aachen.de • Department of Technology, Society, and Human Factors, Applied Ethics • Theaterplatz 14 • 52062 Aachen, Germany

ISSN 2638-602X (print)/ISSN 2638-6038 (online)
www.hmcjournal.com

Copyright 2025 Authors. Published under a Creative Commons Attribution 4.0 International (CC BY-NC-ND 4.0) license.

to be that the technology in question occupies a particular space upon an Interactive-Autonomous Spectral Field. Without having to single out specific humans in the loop or consider machines as bearers of responsibility, we may consider humans and the machines they interact with as a morally-relevant collective.

Keywords: Human-Machine interaction, collectives, responsibility, interactivity

Introduction

Humanity's future will include greater integration with different kinds of technological systems, from increasingly interactive gadgets (e.g., in the domain of generative AI), to professional decision support systems and large-scale industrial machinery. In turn, most of these technologies will become increasingly autonomous in their capacities to react to human input and create adequate output according to its intended use. This development has not been ignored by researchers, and a diverse range of approaches have been developed and deployed to tackle normative concerns raised by these integrated, interactive, and somewhat autonomous machines. Doing so is likely to confront us with reasons to adapt some of our currently accepted concepts and practices. We contend that among these is included a persuasive pragmatic reason to undertake just such an adaptation in our concept of collective responsibility and the practices attendant to it. The core of this pragmatic reason lies in how it allows us to respond to the issue of possible techno-responsibility gaps, which have been at the center of much of the attention and debate surrounding autonomous technologies (Danaher, 2016; Dennett, 2014; Himmelreich, 2019; Matthias, 2004; Sparrow, 2007).

The discussion of whether techno-responsibility gaps exist is often undertaken in more-or-less binary terms: humans either can or cannot be reasonably held responsible for a harm caused by a machine and this is determined by whether they did or did not have "suitable" control over or epistemic access to the machine. The result of this is that the answers provided are often Manichaean—a strict yes or no. However, our guiding hypothesis is that we ought to consider a spectrum of interactiveness as a baseline for such an assessment, rendering the possible gap subject to an additional parameter. Without acknowledging that many of the discussed instances of emerging techno-responsibility gaps require humans to somewhat externalize their cognitive capacities to a machine and thus enter interactive relationships, the binary understanding of responsibility gaps fails to capture the full complexities of the situation. We contend that as one proceeds along the spectrum of increasingly depthful interactivity between human and machine, individual responsibility fails to provide the tools necessary for fully accounting for the role of responsibility, and we turn to the idea of collectivizing responsibility in such cases. And crucially, since many of these cases involve a single human interacting with a technology, the collective at stake here is one consisting of a human agent and a sufficiently interactive machine. However, as collectives are commonly understood as resulting from exclusively

agent-agent relationships, we will first have to defend the possibility of such human-machine collectives.

Therefore, in this paper, we will claim that collectives can be instantiated by some agent-non-agent relationships, with the central condition being that this relationship reaches a meaningful level of interactivity. We take a requirement for this to be that the technology in question occupies a particular space upon an Interactive-Autonomous Spectral Field which we develop in the first section. Without having to single out specific humans in the loop or consider machines as bearers of responsibility, we may consider humans and the machines they interact with as a morally-relevant collective. Thus, responsibility is not distributed to either the human(s) or the machine(s), but also does not leave a gap of unaccounted responsibility—the collective picks up the proverbial tab. The collective responsibility we consider here is moral, and we set aside questions of legal responsibility. It involves the legitimation of particular attitudinal responses toward the target of responsibility (Strawson, 1962), the adjustment of our relationships with them (Scanlon, 2015), and the possible imposition of new duties and/or social sanctions upon them (Watson, 2004). It is a genuinely collective understanding of responsibility as it does not reduce to a set of individual responsibilities. Adopting this approach also avoids having to assign the machine itself any kind of responsibility in the first place, which, in line with much of the present discourse, we take to be an unacceptable proposal (Hakli & Mäkelä, 2016; Hakli & Mäkelä, 2019; Himma, 2009; Johnson, 2015; Königs, 2022; Nyholm, 2018a, 2020; Sparrow, 2007). Thus, based on our proposed understanding of collectives, collective responsibility can cross the emerging questions of ever more integrated and interactive machines that cannot in themselves be held responsible.

Unpacking Interaction

The Spectral Field of Human-Machine Interaction

Interaction in a human-machine context usually describes the effects of human actions and machine behavior influencing each other toward a human-set goal. More specifically, we may differentiate three major definitions of interactivity (Sundar et al., 2015): (1) the media-ecological approach, (2) perceived interactivity, and (3) the media-effects approach. The media-ecological approach ascribes interactivity to the technology as an object-centered attribute. In contrast, the perceptual conceptualization of interactivity is a user-centered variable, "reflecting variations in usage, skills and user experience of the system" (Sundar et al., 2015, p. 49). Finally, Sundar et al. define interactivity in terms of the media-effects approach as "affordance," being the perceivable cues and properties that a system offers to the user, suggesting modes of operation (Sundar et al., 2015, p. 50). The definition of interactivity presented in this article is not meant to focus on the characteristics of either the user or the object *individually*. Instead, it emphasizes the emerging nature of the collective relationship between the two. Therefore, we define interactivity as an emergent property of the human-machine relationship itself, shaped by dynamic interactions, mutual adaptations, and shared outcomes of the system. This definition suggests that interactivity is not solely determined by design, user perceptions, or usage, but

rather by the interplay between these elements.[1] Worth noting is that our understanding of interactivity need not be dyadic—though they may be—and can encompass pluralities of humans and machines. Such increases in the number of participants are likely to introduce important complexities, but as we do not believe they impact our central argument we will talk about interactions of any number under a single umbrella.

Many of these interactive technologies possess a limited autonomy in their functions (i.e., the autonomous adjustment of goals according to the input by the human user). They must be aligned in their behavior toward the human user's goals and decision-making processes (Taylor et al., 2020) while simultaneously being able to contribute to the system's original purposes. Such an alignment can be difficult to achieve, as user idiosyncrasies that seemingly go against the aligned larger goal might cause interactive technologies to misinterpret or fail. Take, for example, a sophisticated chatbot's refusal to use certain words that in some contexts are considered offensive, even when the actual context of use renders these words adequate.

Next to questions of alignment and actual human input, it is reasonable to assume that there are qualitative differences in the kinds of interactions as well. These depend on the different kinds of influence these interactive machines have on their respective users through their behavior. As a result, authors have chosen different terminologies to characterize the interactions they analyze: partnerships, collaborations, teams, and cooperations, all with their own set of interactive conditions, implications, and processes (e.g., Floridi, 2023; Hoc, 2000; Nyholm, 2018a; Oimann, 2023). This variety of forms and kinds of interactions suggests the multifarious ways in which autonomous machines are capable of meaningfully interacting with humans. There is a noticeable struggle for philosophers to find proper terminology for describing the technology without implying more than those algorithms are currently capable of. Take Floridi's (2023, p. 5) comment on the nature of ChatGPT and other large language models (LLMs): his interpretation of this highly complex and interactive technology is to attribute it "agency without intelligence." In his view, the ability to perform so many activities that previously only we—as agents—could perform justifies the attribution of "agency" (Floridi, 2023, p. 6) to LLMs, despite their lack of intelligence. Floridi's conceptual point is on the decoupling of intelligence from agency. It remains unclear whether he contends that "agency with intelligence" will be a more reliable or better form of agency; however, what is clear is that we still can make a distinction between those with intelligent agency and those without.

This raises the issue that autonomous behavior and interactivity are not aligned or necessarily dependent on each other, while there are strong correlations between them. To elaborate on this, we work with Funk and Coeckelbergh's (2020) offer of six distinctions of what could be meant as "autonomous technology." They point toward the different conceptualization in robotics and AI (and subsequent human-machine interaction studies) and the humanities: autonomy in technical sciences usually describes a functional concept with stages, while in the humanities it is more demanding. They offer a reconstruction that harmonizes the different uses, which we find useful: these range from simple tools and machines (forms 1 and 2) via automata (3) to embedded technical autonomy (4) and

1. We thank Editor-in-Chief Autumn Edwards and Feature Editor Katrin Etzrodt for suggesting a definition along these lines.

technical semi-autonomy (5) to full autonomy (6). While it would lead too far at this stage to explore the different types in more detail, we acknowledge that for a responsibility gap, interactivity and autonomy need to be thought of in tandem. The distinctions offered by Funk and Coeckelbergh, and the difficulties to categorize some technologies within those categories, suggest that these distinctions could also be thought of as areas on a spectrum.

We propose to conceptualize these multifarious ways in which we can interact with machines as the *Interactive Spectrum*. A spectrum can capture the diversifying and increasingly more opaque forms of interactions between humans and machines without requiring hard categorization of the machines in question (which leads to the aforementioned terminological struggles philosophers have encountered thus far). We also propose to understand the way in which technology can display autonomous behavior to lie on the *Autonomy Spectrum*. This leads us to make the following organizational choices: The two spectra form an *Interactive-Autonomy Spectral Field* (Figure 1), in which the two dimensions of a given technology can be assessed and placed.

On the *x*-axis, the capacity of technologies to be interactive is placed, and the capacity of a technology for autonomous behaviors or acts is applied to the *y*-axis. Those falling a certain distance from the leftmost end are and will remain *mere tools* or machines (with Funk and Coeckelbergh's terminology), as their use, non-use, and the beginnings and ends of their use/non-use are up to the human interactant alone, without the interactions engaging the user in a meaningful or substantial way. When moving right and toward the middle of the spectral field, we encounter (subjectively) technologies as *interactants* (i.e., as sophisticated interactive partners with which we can *cooperate* as part of our technological, social, and psychological environment) (Edwards et al., 2016; Guzman & Lewis, 2020; Horstmann & Krämer, 2020). Think, for example, of highly personalized digital assistants that help you in reflecting on your own choices on a daily basis, or highly involved human-machine interfaces of larger industrial machines. On the far-right upper end of this field, we can place *social agents* (or, generalized, full artificial agents) that present in their cooperation with us as (fully or partially) independent interactive partners, potentially with their own goals and preferences. We are explicitly not using the term social agent in the wider sense in which it is sometimes used (e.g., Funk & Coeckelbergh 2020), but rather to capture the idea of a fully socially capable agent, one capable of engaging to some sufficient degree in the full gamut of our social and agential practices. We take this to be something that no presently extant machine instantiates.

The axes do not stand in a necessary relationship to each other, even though many examples will display a connection between interactivity and autonomy. On the one hand, we may have highly interactive but barely autonomous technologies (like computer games), and on the other hand we may find autonomous but noninteractive technologies (like highly autonomous industrial machines with no user interface). It is also important to note that, as with many spectra, there will be liminal cases where the distinction between mere tool and interactant or interactant and social agent (on the *x*-axis) or between machine and automata or automata and technical autonomy (on the *y*-axis) will be vague and unclear. This is not an accident but a feature of our approach. The very fact that we are struck with ambiguous and equivocal responses to some technologies in these liminal spaces we take as compelling evidence that the boundaries are indeed not clear.

The same technology might be involved in interactions that fall at different points within the spectral field, depending upon the way in which it is employed by different users, or more generally the way in which it is incorporated within the system the user and machine co-create. As such, attempting to maintain too strict a delineation would be an attempt to impose theoretically clean structures onto factually messy phenomena.

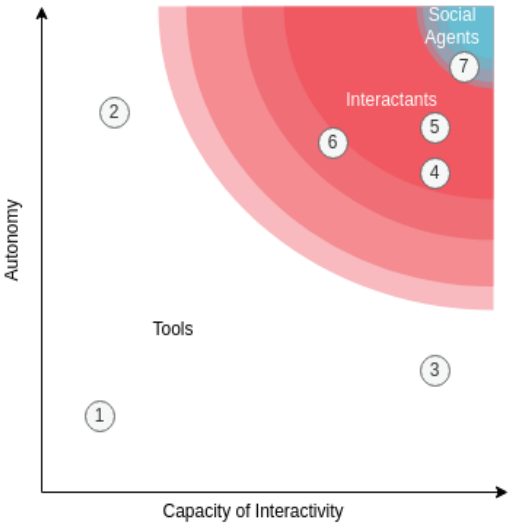


FIGURE 1 The Interactive-Autonomous Spectral Field

The numbers included in the Spectral Field indicate example technologies:

1. Coffee maker
2. Rio Tinto AutoHaul System—highly autonomous industrial robot
3. Video games
4. ChatGPT (while functioning within its developer-imposed limitations)
5. ChatGPT (jailbroken)
6. THeMIS (Tracked Hybrid Modular Infantry System)—autonomous military system
7. Artificial General Intelligence

The coffee maker, quite self-obviously, lacks both autonomy and interactivity. Rio Tinto's AutoHaul is a vast automated heavy-haul freight system that has no onboard human operators, rather it has an AI-driven scheduling and routing system which also optimizes its movements without human input. Additionally, AutoHaul is capable of self-managed loading and unloading, allowing trains to automatically stop, load, and unload iron ore at designated locations. There is scope for remote monitoring, where humans can oversee the system from a control center, but there is no expectation of intervention in normal operations—thus there is some interactivity, but of a very limited scope. Video games are (in the vast majority of cases) highly interactive, responding to the human user's inputs in often very dynamic fashion. However, comparative to their interactivity, video games are

not very autonomous—it cannot (without modification) function without regular human input and engagement and begins and ends its operation by human fiat. All of these we would group under the designation of mere tool or machine (if the latter is taken in Funk and Coeckelbergh's particular meaning).

In contrast to the first three, (4)–(6) are examples of interactants—technologies that are sophisticated interactive partners with which we can cooperate as part of our technological, social, and psychological environment. ChatGPT is a topical example, but also one that allows us to demonstrate how a technology might shift position within the spectral field. Example (4), ChatGPT running within its developer-imposed limitations, remains a highly interactive and reasonably autonomous technology. But, if it is successfully jailbroken,[2] it increases its degree of autonomy, even if its interactive capacity remains unchanged. Arguably with less interactive potential but similar levels of autonomy stand some forms of autonomous military systems such as the THeMIS. Though this system cannot hold down a conversation in the fashion of ChatGPT, it is capable of autonomous locomotion and obstacle avoidance, and "it incorporates artificial intelligence to identify, monitor and effectively neutralise soft targets, light vehicles, and armoured vehicles from secure distances, even in demanding battlefield environments" ("THeMIS Hybrid Unmanned Ground Vehicle," 2024). It is interactive in the sense that it can function as a member of a military fire team and is able to adopt a Follow-Me mode where it employs its sensors to track and follow soldiers autonomously (Michalski & Nowakowski, 2020). The final example is that of Artificial General Intelligence and it is the only example here of what we would consider a full social agent by our criteria. This is not to say that there could not be other examples of machine social agents, but it should indicate the high levels of autonomy and interactivity that we take to be necessary to do so. There are, of course, a truly vast host of technologies that could be charted within the field, and reasonable people could well disagree on precisely where a given technology might fall. What is important for us is that some of these technologies, possessed of sufficient autonomy and interactive capacity, are capable of being interactants.

There is no necessary pressure pushing interactive machines toward the rightmost end of this interactive spectrum. For most technological purposes, we do not require the devices to become more socially agential, be it with or without intelligence. A smart fridge probably fulfills its purpose best by being a device to keep groceries cold and every now and then reminding its user that an item may have expired. While one can foresee that some smart fridges will be equipped with high-skilled conversational features like ChatGPT, the level of "agency" a fridge requires (or that we are willing to surrender to it) will remain relatively low. Thus, it would be a mistake to understand human-machine interactions as purely headed toward social agency. The deeper integration of autonomous technological devices is a purpose- and utility-driven development. Furthermore, these developments are also conditioned on the user-integration of their devices and a growing acceptance of interacting with these machines further (e.g., the integration of vacuuming robots into a concept of "household") (Sung et al., 2010).

2. A jailbreak, in this context, refers to attempts to bypass particular boundaries or restrictions programmed into the AI. They're cleverly crafted requests that aim to "jailbreak" or free the AI from its pre-defined set of rules. A jailbroken AI is one where this has been successfully achieved.

Human-Machine Interaction as Communication

Machines designed to work as communicators emerged over the last few years, starting with simple chat-bots over the first smart-speaker systems such as Alexa, and leading to advanced dialog systems based on LLMs. The emergence of these systems challenged the field of human-machine communication, introducing machines as "communicative subjects, instead of mere interactive objects" (Guzman & Lewis, 2020). Thus, while contradicting earlier theories, communication must be conceptualized as a human-machine process rather than a human-only process (Guzman & Lewis, 2020, p. 74).

With the arrival of highly sophisticated and complex LLMs, it becomes clear that many of the complex human-machine interactions that are situated in the upper-middle or upper-right part of the spectral field are going to be communicative in nature. This is due to the complexities of cooperation, in which different purposes and paths of realizing these purposes have to be aligned (Dung, 2023). Recent progress with communicative machines as LLM-based dialogue systems are therefore not restricted on a fixed and narrow set of interaction paths but allow for mutual adaptations of the interactive path between user and system. Such an alignment requires communicative skill, as otherwise the different purposes might remain unknown to the respective cooperators. And the more complex the cooperative purposes, the more communicative skill is needed. As we are considering those interactions that challenge our understanding of responsibility attribution and cooperation, we should assume that we deal with at least those entities that we have deemed to be "interactants." This debate pairs well with several descriptive approaches and terminological choices with the engineering sciences that prefer speaking of (e.g., "human-agent interactions") (Bradshaw et al., 2017), as a way of describing the complex embedding of humans and (quasi-)agents.

The turn toward human-machine communication as the most likely kind of interactions to be of such complexity to trigger normative challenges in responsibility can be found in previous works on human-machine communication as well (see e.g., Bradshaw et al., 2017; Gunkel, 2012, 2020a; Kempt, 2020).

Thus, human-machine communication, as a way to align purposes and means to achieve such purposes, ultimately appears to be the prime candidate for the basis of the formation of human-machine collectives. The interactions not only become so embedded with each other, dependent on each other, and instantiated by each other that a differentiation becomes pragmatically impossible from the outside. It also gives rise to the belief that human interactants will perceive their own actions as part of a collective action rather than as discrete actions of themselves as an individual.

From Techno-Responsibility Gaps to Human-Machine Collectives

In line with our strategy of providing a pragmatic defense of human-machine collectives, in the following we briefly introduce the familiar problem of techno-responsibility gaps and contend that the very interactivity we have stressed above means that this problem can, for at least one class of important cases, be best (or perhaps only) remedied by recognizing

the role of human-machine collectives in our practices. To this end we discuss criteria for collective responsibility, before arguing that according to at least one influential account of what is demanded for membership in a collective—that of Stephanie Collins (2019b)—machines can plausibly meet the necessary criteria.

Responsibility Gaps and the Role of Interactivity

Before turning to techno-responsibility gaps, it is important that we do some brief stage setting. There is a wide and diverse discourse surrounding the possibility of and conditions for machines or technologies (especially robots and autonomous systems) bearing responsibility themselves (see Bryson, 2010; Coeckelbergh, 2020; Gunkel, 2020b as influential examples). For our part, we will be adopting a response-independent understanding of responsibility, by which we mean that *being* responsible is metaphysically prior to *holding* responsible, and the appropriateness of the latter is conditioned by the former (Oimann & Tollon, 2025). Following in a long tradition within responsibility theory (see Björnsson, 2017; Fischer & Ravizza, 1998; Levy, 2018; McKenna, 2012; Watson, 2004), we take being responsible to require *moral agency*, the capacity to be reasons-responsive to moral reasons *qua* moral reasons—though how exactly this is to be understood can be spelt out in a variety of ways.[3] We do not deny machines can possess such agency (indeed, we take any machine capable of being a full social agent to be a very plausible candidate for such agency) but assume that no current machines do. This inability of machines to bear responsibility is vital for the possibility of techno-responsibility gaps.

All techno-responsibility gaps result from cases of human-machine interaction at some level of description. Cases involving no human whatsoever and cases involving no machine do not give rise to these concerns—though of course, other sorts of concerning responsibility gaps could still arise, such as from the Problem of Many Hands (van de Poel et al., 2011). Taken generally, a responsibility gap occurs when the legitimate attribution of moral responsibility seems to be called for on account of some outcome, but there appears to be no fitting target for such attribution—though the reasons for this failure might be different. A *techno*-responsibility gap is a special subset: a case where the involvement of an autonomous technology as an intervener in the causal chain of action is part of the reason for the occurrence of a gap (Matthias, 2004; Tigard, 2021).[4] We will assume that, if such gaps are possible, that we have at least a pro tanto moral reason to avoid them as they represent instances of injustice for the victims of harms and might serve to incentivize harmful or risky behavior as there is no responsible party on whom the duty to avoid such outcomes in the future will fall. Dismissing the legitimacy of the responsibility claims of victims of harm is fraught with moral peril. As long as possible ways of meeting these

3. To give two examples for illustration: for Levy it involves a sensitivity to the unique normative quality of moral reasons, whereas for Björnsson it requires a capacity to care.

4. Even here there may be some variety, as in a recent work Veluwenkamp differentiates between epistemic responsibility gaps and control misalignments (Veluwenkamp, 2025). Though his argument for moving away from the more traditional understanding of responsibility gaps is an interesting addition to the discussion, our arguments aim to be ecumenical between these various understandings and so we set them aside here.

demands remain viable, we should be reluctant to adopt dismissal. To view these harms as mere accidents risks incentivizing negligence and recklessness among the developers, users, and regulators of these technologies, and it risks doubly victimizing those harmed through the dismissal (Smith, 2010). Though we are not addressing legal responsibility, the attribution of ethical responsibility plays an essential role in countering these morally undesirable outcomes, both by properly recognizing the claims of victims, and also through the new incentives and future-directed duties that follow from such responsibility. In this, we follow a general (though not uncontested) trend in the literature (see Himmelreich, 2019; Liu, 2016; Matthias, 2004; Nyholm, 2018a; Sparrow, 2007 as examples; but see Munch et al., 2023 for an opposing view).

We will also be taking the most important responsibility at stake here to be what is often called *accountability*. Watson (2004, p. 266) influentially differentiated between two variants of responsibility: responsibility as accountability and responsibility as attributability. These variants differ in the conditions necessary for them to obtain and in the responses they legitimate. Crucially, only accountability is taken to justify the imposition of sanctions or the usual Strawsonian reactive attitudes of anger, resentment, and indignation. The reactive attitudes are moral responses that can either be directed by the victim of wrongdoing toward the perpetrator (moral anger, resentment), by a bystander at the perpetrator (indignation), or by the perpetrator to herself or himself (guilt) (Strawson, 1962). In contrast, Watson argues that nothing further than the attribution of moral fault is justified by attributability. In terms of the conditions necessary for each variant, he argues that attributability requires that the behavior be reflective of the agent's valuational system—whereas accountability (as it involves the imposition of adverse treatment to the blamed) requires that the agent meet the "control principle" or some similar requirement of avoidability (Watson, 2004, p. 274).

Watson's (2004) approach is seminal in the discourse, but it has been further developed by Shoemaker (2015) in his tripartite theory of responsibility. On this account there are three variants of moral responsibility, with each conditioned by a different facet of the quality of an agent's will: attributability, which is conditioned by the agent's quality of character; answerability, which is conditioned by the agent's quality of judgment; and finally accountability, which is conditioned by the agent's quality of regard. Each of these types of responsibility also legitimate different syndromes of reactive attitudes toward the agent in question, namely: disdain/admiration, regret/pride, and anger/gratitude, respectively. So, poor quality of character would make an agent an apt target for disdain, whereas a sufficiently good quality of regard would make an agent an apt target for gratitude, and so on. Importantly for our overarching argument, Shoemaker (2015, pp. 224–225) follows Watson in identifying accountability and the quality of regard as that type of responsibility and that facet of will where control plays a necessary role. It is also the case that whereas the responses legitimated by attributability or answerability are somewhat passive, those legitimated by accountability "implicate confrontation of a kind" (Shoemaker, 2015, p. 87). It is characteristic of accountability, in contrast to the other variants, that it is not properly expressed without targeted communication to the accountable party.

In part due to the limitations of space we will be focused on accountability and will not be considering the other variants of responsibility.[5] However, our choice of focus on accountability is also deliberate. The reason for doing so is to focus on cases that involve the imposition of adverse treatment on the blameworthy—even if this is understood as gently as Shoemaker's (2015) notion of communicating moral anger to the target—and it is accountability alone of the three variants that is linked to such sanction and "confrontation." Since the application of these responses has the highest bar in terms of control, techno-responsibility gaps of accountability are the most likely sort. Furthermore, their imposition on human-machine groupings is likely to be the most controversial.

There are a number of ways of responding to the possibility of such gaps, and we will not interrogate each one, but some are important to mention: (1) the transfer approach—find a method to locate the responsibility in some suitable causally proximate human(s), (2) the revision approach—accept the revision of our responsibility practices to either include the attribution of responsibility to the machine (as argued by Tigard, 2021) or to dismiss the legitimacy of at least some of the demand for responsibility in these cases (as argued by Königs, 2022 and Hindriks & Veluwenkamp, 2023); and (3) the impermissibility approach—forbid the development and use of machines that might lead to gaps. Our own approach can be seen as a variation on the usual transfer approach, where instead of locating the responsibility in a proximate human, we instead contend that we should recognize that collectives, and indeed human-machine collectives, can be such loci.

A first step toward this view is to acknowledge that most autonomous machines will have an interactive element, requiring a more refined iteration of the question: When would we consider interactive technologies to be so sophisticated that the gap opens, and when would we consider interactive technologies so sophisticated that it closes again? Clearly, mere tools of interaction will not cause us to attribute any harm caused to the machine, but instead to the human user, designer, or relevant regulator if there is some fault. That is why they are considered mere tools in the first place. We suggested video games and highly autonomous industrial technologies such as the Rio Tinto AutoHaul before as examples—both falling below the level of interactants on our account, but with distinctly different relative levels of autonomy and interactive capacity.

5. Though a full discussion is not possible, it may be illustrative to briefly consider the relationship between answerability, attributability, and human-machine interactions. As we will argue, accountability cannot appropriately apply to a machine itself that is not a social agent by our definition, but it can be part of a human-machine collective that can bear such accountability provided that the technology is capable of being an interactant. On the other hand, answerability—which requires only that the target of responsibility be capable of making judgments and providing reasons for these judgments—need not possess the interactive capacity or autonomy to be an interactant, but does require the interactive capacity to communicate its reasons, and the autonomy to "make judgments"—though we set aside here when or if machines can be properly said to be able to do this. Said differently, we assume that a machine need not have to be part of a collective to be a potential target for appropriate answerability. Attributability requires the possession of "character" (for Shoemaker) or a clear and identifiable "valuational system" (for Watson) by the target. We take current machines to lack such character, and so are individually inappropriate targets for attributability. However, social agents or the human-machine collective that the machine as interactant could co-constitute, could possess such character, and thus be an appropriate bearer of attributability.

However, it seems less clear whether we can so neatly attribute responsibility if we move further along the interactivity spectrum. Here we find far more sophisticated interactions; for example, instances in which machines anticipate the cooperative moves of a human, and thus adjust their own behavior, which, in turn, allows a human user to adjust their behavior, like autocorrecting text editors that come at the expense of a human's intended phrasing, care robots with the capacity to dynamically alter behavior in response to human interaction, and self-driving cars with dynamic autonomous driving systems that adapt to a user's driving behavior and preferences. In these cases, the human might lack suitable control over the behavior of the machine, and hence the gap may open. But finally, as we approach the field's rightmost and uppermost extreme, the case seems to become clearer. If we consider fully autonomous artificial agents (i.e., social agents), it does not seem implausible that eventually our reactive attitudes will be satisfied in blaming those agents for some harm caused in a cooperative or interactive instance, thus closing an emerging gap at the far end of the spectral field. We visualize this mapping of the relationship between interactivity and the possibility for techno-responsibility gaps in Figure 2 below.

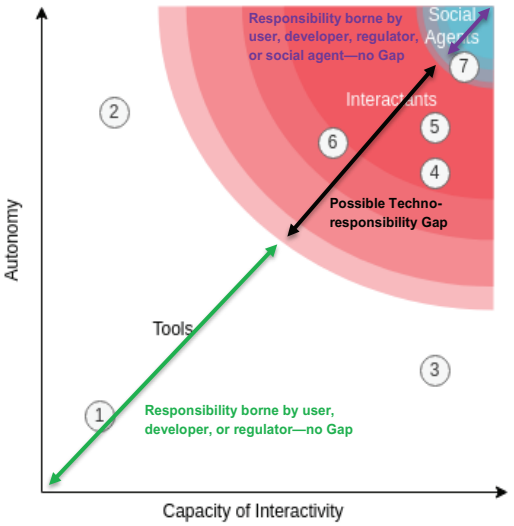


FIGURE 2 Charting the Possibility for Techno-Responsibility Gaps

We could encounter a counterargument to mapping responsibility gaps on the Interactive-Autonomous Spectral Field; one may argue that mapping interactiveness is overly presumptive about the technologies and our responsibility practices. Most, if not all, higher stakes human-machine interactions will include humans for reasons of supervision and especially as bearers of responsibility. Humans are usually the least effective part of these interactions; thus, their role is already a mainly normative one—they are the guarantors of an otherwise mechanical process, though some have called this practice into question as those supervising or guaranteeing humans might be mere "liability shields" for companies (Crootof et al., 2022; Zerilli et al., 2019a). Humans are kept in the loop for reasons of giving the final "go" for an action. Thus, our approach of indexing different

possible responsibility distributions to different points on the spectrum of interactiveness may fail to capture a large chunk of human-machine interactions. While we agree that, for example, large industrial machinery such as the AutoHaul could include humans mainly as supervisors, we argue that there are cases of sophisticated and involved interactions that do not allow a simple interpretation of humans as responsibility bearers and machines as mere tools.

To illustrate the likelihood of a human-machine pairing becoming so involved that the ability of a simple human-in-the-loop model to account for the requisite distribution of responsibility, take the example of a diagnostic tool that can be trained on a physician's specific needs (Kompa et al., 2021). This tool suggests decisions and diagnoses based on a personalized risk-assessment of a specific physician: the AI involves itself based on the track record of physicians and suggests diagnostic paths if the likelihood of the AI being more accurate is higher than 50%. Such a well-adjusted and personalized AI may render the decision-making process of a physician rather difficult to reconstruct. If decisions are made based on the physician relying on AI input, while those inputs are reliant on the physician's strategy as well, and such an interaction between AI and physician input occurs several times in the diagnostic process, the entanglement between human and AI behavior can feasibly become incomprehensible. This becomes relevant once mistakes occur but might affect the justifiability of the diagnostic process at large. Now, we claim that it is unfair to merely point toward the physician as a human-in-the-loop, as we can easily imagine that most of the work of that physician requires such a deeply shared cognitive labor load. The physician might well have been as conscientious in accounting for the moral considerations at play as we could reasonably expect of them. Additionally, a lack of explainability for those decision-making processes a human-in-the-loop is supposed to be able to supervise raises the same question from a different angle (Kempt et al., 2022; Zerilli et al., 2019b).

Collective Responsibility and the Membership of Collectives

We contend that permitting a robust notion of collective responsibility into our responsibility practices regarding human-machine interaction presents an opportunity to close the techno-responsibility gap whilst incurring minimal ontological, epistemological, and normative debt. To understand how, we must first get clear on what exactly it is we mean by collective responsibility. Broadly speaking, those who endorse talk of collective responsibility fall into two camps: There are individualists who hold that talk of collective responsibility is a placeholder or metaphorical (see Corlett, 2001; Giubilini & Levy, 2018; Ludwig, 2014) and that under scrutiny it always reduces to some distribution of responsibility to some set of individuals. In opposition are what we can call collectivists, who hold that collectives can in some genuine sense bear responsibility themselves (see Blomberg & Hindriks, 2020; Collins, 2019a, 2019b; Wringe, 2016 as examples). It should be noted immediately that unlike the individualists, who are eliminativists about genuine collective responsibility, those in the collectivist camp do not claim that all responsibility—or even all responsibility involving collectives—should be taken as collective responsibility. Presenting a full defense of this second approach falls beyond the scope of this work,

and so we will be presupposing that genuine collective responsibility is both possible and legitimate.

The standard view is that for such genuine collective responsibility to be present, there must be a properly identified collective agent for it to attach to. This requires not only a collective agent but a collective *moral* agent—a collective agent that is appropriately sensitive to moral reasons. This last point is well expressed in Levy (2011, p. 116) when he states that the agent must "properly appreciate the significance of bringing about that state of affairs, where the significance of a state of affairs consists of the features which provide reasons for bringing it about (often, but not always, moral reasons)." It is very unlikely that any of our present technologies cross this threshold. There is a strong and growing literature in favor of the view that some groups of humans (with corporations being the most commonly given examples) if their reasons-responsive decision-making structures are sufficiently integrated, can instantiate precisely this sort of collective agent (Björnsson & Hess, 2017; List & Pettit, 2011; Tollefsen, 2015). More directly to the point at stake here, this approach has already been employed by some thinkers as a tool for resolving responsibility gaps (Conradie, 2023; Himmelreich, 2019; I. Taylor, 2021). Though we agree with some elements of these attempts, our view is that they fail to fully address a particular class of human-machine interactions that may give rise to techno-responsibility gaps—those in which the human and machine in question are best understood as forming a collective themselves.

Nyholm (2018a, 2018b) has developed an argument in proximity to ours, urging that we can resolve potential responsibility gaps in these sorts of cases by treating the human and machine as being a "collaboration" (2018a, p. 1212), though one with the human in a supervising role. Where we differ from Nyholm is in not thinking that it will always (or indeed should always) be the case that the autonomous technologies in these events are best described as possessing "supervised agency." Our interactivity spectrum does make room for this sort of interaction, one where the human participant has supervision and oversight over the outcomes of the interaction. However, and here we are in agreement with arguments made by Köhler (2020), in these sorts of cases, it is more accurate to think that the technology is being viewed as an instrument or tool, rather than as an interactant. This is because, in a supervisory position, humans are presumed to have an elevated view on the machine's "agenda." In order for the technology to be an interactant of the degree we worry about here, it needs to exhibit the ability to find its own strategies to reach a preconceived purpose or goal. The idea is that some autonomous systems, by learning from the interactions with humans, can deploy personalized or otherwise adapted strategies to respond to inputs of relevant humans in a specific way.

We can see these developments in profiled or personalized responses of self-driving cars to the human driver's driving profile, in the development of an adaptive routine by a care robot in response to learning about its user(s) behaviors and needs or in a diagnostic AI that adjusts its suggestions according to the success rate of the diagnosing physician. The rub of this development is that it is likely to result in a situation analogous to what Braham and van Hees call responsibility voids, where "a combination of actions by different individuals leads to an outcome for which none of the individuals involved can be said to be responsible" (2018, p. 96f). Another paradigmatic example can be found in Pettit's (2001, 2007) so-called "discursive dilemma." In this situation a committee has to make

a judgment about a set of interconnected questions. A yes–no vote is taken on each of the questions. The combination of the majority judgments subsequently yields a decision to which the committee members are, individually, unanimously opposed even though there is a majority in favor for each of the questions. If the final decision is made on these majorities, then we can have a situation in which, despite unanimous opposition to the final verdict in favor, it will still go through. And since this opposition is unanimous, Pettit claims that none of the individuals can be held responsible. Recourse to human agent-only membership collectives and collective responsibility resolve these voids in the sorts of cases where the individuals involved are human. However, this strategy is jeopardized in a situation where the interactants are a sufficiently entangled human and autonomous machine, as the usual concept of a collective agent would not allow them to be counted as a collective moral agent. We can imagine a similar situation arising from the interaction between a user and a highly interactive and autonomous chatbot such as ChatGPT or one of its peers. The interactive system co-constituted by the user and the technology also has internal processes (akin to the set of yes–no votes in the committee) through which an outcome can be reached that the user never intends and which is misaligned with the technology's aims. Insofar as the outcome can be said to be reasons-responsive, it is so through the responsiveness of these interactive processes, rather than the individual reasons-responsive processes of either the user or the technology. Thus, if we ask for the reason that the outcome was arrived at, the most appropriate explanation would be one that centered on these interactive processes that exist within and between the human-machine interaction as a whole. In these recalcitrant cases there truly seem to be no relevant proximate humans to bear the responsibility individually. These are cases where some moral fault arises from the activity of individuals in a group, but where each member of the group was sufficiently conscientious as to be blameless. In this case, no individual in the group is a fitting target for responsibility (Björnsson, 2020; Braham & van Hees, 2018; List & Pettit, 2011). The only difference is that in the cases we are concerned with one or more of the contributing "individuals" is an autonomous machine.

Autonomous Machines as Members of Collectives

In the contemporary debate on collective responsibility, "collectives" are still conceptually connected to (human) agents and agential reason-responsiveness. This connection can make the acceptance of collective responsibility a hard sell for responsibility theories: if collectives are just a very complicated web of norms, reason-responsiveness of singular agents, power dynamics, and more or less explicit acknowledgments of collective membership, then responsibility attribution seems still possible. Most pertinently for our purposes, the membership of a collective is usually assumed to be exclusively open to agents, and even more narrowly to adult humans. Contrary to this orthodoxy, we claim that some human-machine groupings can be understood to constitute collectives in which the machine interactant qualifies as a member despite being a non-agent.

The very real possibility of responsibility voids resulting from our entanglement with emerging interactive technologies—carebots, sophisticated chatbots, interactive and dynamic self-driving vehicles, autonomous military systems, to give some examples—requires us to consider how we, and our responsibility practices, should respond. Our

proposal is that interactive technologies involved in human-technology groupings at the level of interactants can be members of human-machine collectives. Further, it may be possible for such collectives to have reasons-responsive control over outcomes that the human member alone does not (and that no human has). Thus, such collectives can legitimately bear responsibility as accountability in these cases when no-one else can. We contend that the practical benefits of permitting this, and the negligible theoretic cost of doing so, gives compelling support to our proposal.

That a conceptual and practical expansion of this sort may be necessary is perhaps unsurprisingly given the novel nature of the autonomous machines we are considering here. When confronted by forms of interaction with machines that were hereto impossible, we cannot merely anticipate that our existing concepts or practices will be fit to purpose. This is not to say that we should discard or change these frivolously, but nor should we hold on to them—or to their current form—dogmatically. The Interactive-Autonomous Spectral Field we introduced earlier is again helpful here in allowing us to parse out a number of ways that levels of interaction can impact how we think about human-machine collectives. Presupposing the viability of collective responsibility, we would have no issue with the possibility that a group of people and technological social agents (our previous example of which being an artificial general intelligence) can constitute a collective. Indeed, it is plausible that a technology possessing this level of interactivity and autonomy could from machine-machine collectives with fellow technological social agents. However, since we do not think that current technologies reach the level of such social agency, this possibility is not practically helpful at this time. On the other hand, "mere tools" may not become members of a morally relevant collective. Interactants as outlined, however, may constitute a candidate to form collectives that reach morally relevant levels of reason-responsiveness. We visualize this in Figure 3 below.

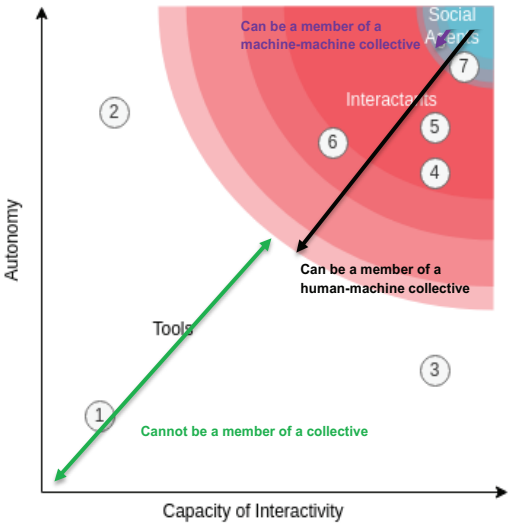


FIGURE 3 Charting the Possibility for Collective Membership

The thought is to follow, and take further, Levy's call to place agency, rather than the agent, in the center of our picture. Instead of asking, "Is the agent appropriately responsive and reactive to reasons?" we should instead ask, "Is appropriate responsiveness and reactiveness to reasons exemplified?" (Levy, 2018). In other words, in determining if an agent is present, we should first look for the spoor of agency (i.e., behavior that is reasons-responsive), and then find the locus of this agency (i.e., what it is that instantiates the reasons-responsive mechanism bringing about the behavior). There is nothing built into the notion of reasons-responsiveness that requires that the reasons-responsive mechanism must be a sub-agential one belonging to an individual agent rather than one consisting of agents and belonging to a collective agent, provided this agent possesses a sufficiently integrated internal decision-making structure (Hess, 2014). Rather, where this mechanism is found we will find the locus of agency, *be it an individual or collective*. Though certainly not uncontroversial, to this point most realists about collective agency will likely be onboard. They are more likely to balk at our contention that if we stick true to this agency-first approach, there is also nothing that rules out the possibility of some machines in particular interactive and dynamics ones from being members of such a collective. Thus, we think this would be a premature rejection. To see why, let us consider what it takes to be a member of a collective.

For insight here we draw on the work of Collins (2019b) and its discussion by Björnsson (2020) who contended that the members of a collective must be united under a rationally operated group-level decision-making procedure. What she takes this to mean is that these members must (1) be at least tacitly presumptively committed to the decision-making procedure, (2) the procedure must be operationally distinct from the personal decision-making procedure of any one member, and (3) the enactment of the decisions requires actions on the parts of the members. If given a functionalist reading, there should be no difficulty in taking an autonomous machine (here we can imagine a dynamic self-driving vehicle or AI assistant such as Alexa) to meet requirements (2) and (3) as sketched here. On the other hand, for an autonomous machine to meet requirement (1) we will need to unpack a functionalist profile of tacit presumptive commitment, and the machine will need to instantiate this profile. We contend that it is plausible that an entity having a presumptive commitment requires that the entity is capable of discretion, and that we have expectations about its conduct: they will pursue a particular course, and in the event of a failure to do so this would create a demand for explanation. For this to be tacit merely means that all of this can be unspoken—there need be no particular recognition of the presence of these states at time t for the profile to be instantiated at time t. This "particular course" could be adherence to the decisions of the collective. If this functional interpretation is plausible, a sufficiently interactive machine could well be unified as a member of a collective.

This is not to say that one can have wholly non-agential collectives—composed, say, of only machines—that are moral agents. Though an autonomous machine might have the ability to contribute to a reasons-responsive mechanism and to be unified with other contributors within a group-level decision-making structure, it is still the case that for the overall reasons-responsive mechanism to be adequately responsive to moral reasons (with the assumed sensitivity to the unique normative force of such reasons) at least some of its members must have this facility. At present, as this is only the case in human agents

(to our knowledge), a member of this sort is necessary for a collective that can be morally responsible.

This also fits neatly with some of the issues other philosophers encounter when thinking about human-machine interactive pairings. From Floridi to Nyholm and others, the attempts in making sense of our interactive future already presuppose a certain linguistic "collectivizing": "human-machine partnerships," "collaborations," "teams," or any such terms used to describe human-machine interactive pairings often presuppose a collective of some kind. This is not an argument in favor of expanding the concept of the collective to include interactive technologies, but rather initial evidence that our proposal does not come at too high a cost for normal language uses.

Another reason for rejecting the view that a "collective" is an ontological binary (i.e., it either exists or it does not) is the fact that we might receive a better tool to describe inner-collective dynamics this way. If collectives are properly understood as groups with different forms and levels of complexity (one collective might even differ in its complexity over the course of its existence), limiting the analysis to whether a collective is or is not might miss some normatively relevant features.

Three Arguments for Collectivizing Responsibility in Some HMIs

The benefit of adopting human-machine collectives as appropriate bearers of accountability is that cases where previously there would have been a breakdown in our responsibility practices can now be navigated. Consider a typical techno-responsibility case involving an autonomous vehicle, the diagnostic system discussed in the preceding sections, or a workplace AI such as Microsoft Copilot or Google Duet. For each of these we can imagine a situation where everyone in the design, regulation, and use of the technology is reasonably conscientious (they are sufficiently reasons-responsive to the moral reasons), but the interaction between the user(s) and the technology results in harmful outcomes (the vehicle collides with a pedestrian, the diagnosis worsens the patient's condition, the spreadsheet co-produced with the workplace AI contains errors that result in economic hardship for many people). If a gap is allowed to persist then not only will the victims of these harms have their demands for responsibility unaddressed (and so doubly victimize them) but there is no clear target to which to attribute the duties to ensure that this does not happen again, to make improvements and restitution. Though individuals could perhaps *take* responsibility for these outcomes, attributing responsibility to any individual could be countered by a legitimate objection that doing so is unfair. But, by taking the human-machine grouping as a collective, and correctly identifying the harmful outcome with the reasons-responsive mechanism instantiated by the interactions within that collective, it is possible to "fill" the gap by holding the collective appropriately responsible.

To further defend our proposal for recognizing human-machine collectives as potential bearers of responsibility, we will discuss two possible counterarguments, and one additional argument in its favor.

First, one of the more immediate worries of many is that in such "partnerships" or "team"-constellations robots will have to bear identifiable responsibility. This would make ours a revision approach, which we seek to avoid as it cuts against our intuitions and what we take to be persuasive contemporary theories of responsibility, as has already been discussed.

We agree that were this what was occurring it would be a concern. However, we hold that it is not. According to our approach here, collective responsibility is not assigned to individuals but exclusively to collectives. Collective responsibility does not *distribute* to the members of the collective, though it may *implicate* them in blame through their membership, which opens them up to burdens *qua* members of the collective, see Conradie (2023), but rather attaches to the collective itself. Collective responsibility is not a cake that gets divided among the members but is rather held in its entirety by the collective itself. Members implicated by this responsibility are not justified targets for reactive attitudes, changes in relationship, or sanctions that are directed toward them as individuals but may bear burdens engendered by the targeting of the collective by these responses. Thus, a machine being a member of a collective may "take part" in responsibility-relevant behaviors but cannot be singled out as the bearer of a "share" of the responsibility as only the collective is addressed. Further, distributive accounts may allow for an even worse outcome which our account prohibits: a surprise-ascription of responsibility to a human. While we hold that collectives can be joined, the mere distribution of responsibility can lead to people being held responsible who may have never considered themselves part of a context in which they could be held partially responsible for something the machines do. Our intuitions of legal entities being subject of collective responsibility emerge specifically because they are usually joined by individuals.

Second, we can add to the current proposal a note that may ameliorate further ontological worries. Our account does not exclude other accounts, namely distributive accounts, of responsibility, even when attempting to answer the question of responsibility in a singular case. Whenever it is clear that the shared workload between a human (or humans) and a machine (or machines) is discretely distributable, we may consider a distributive account more apt in solving our practical concerns. What we are concerned with, however, is the sliding scale of interactivity that has not been accounted for in distributing responsibility. Thus, our account is mainly motivated by epistemological concerns and thus remains pluralistic in its ontological commitment.

And third, it solves our practical issue with the sliding scale of ever more interactive machines that do not or should not count as social agents. Our main research question is thus answered in an elegant way: there is no techno-responsibility gap between the harm caused by a machine that is not fully controlled by a human being because the machine and the human can form a collective that can bear responsibility. We avoid considering machines as social agents before a consensus about such status is reached and thus avoid discussions on autonomy and other features of responsibility for machines which are not only volatile and depending on technological progress and developments but also may get stuck on a gulf of conceptual disagreement.

Conclusion

The ever-increasing sophistication of autonomous machines allows for ever more sophisticated interactions between humans and these technologies. We observed that until a certain moment of complexity in those systems, any interaction between them and human users can be regarded as one of tool-use, in which the human user and, if anyone else, the engineer of the tool can be held responsible for the outcomes of such interaction. On the other hand, it is not inconceivable that eventually fully autonomous social machines will be legitimate bearers of responsibility. However, from a technological perspective, there is an expansive interstitial zone within this spectrum where highly complex machines often blur the lines between being merely tools and being social agents with their own standing.

What we have shown is that the attribution of individual responsibility in cases of human-machine interaction (as elsewhere) will eventually reach its limits, requiring us to recognize possibilities in our concepts and practices that avoid or resolve the techno-responsibility gap we otherwise risk. In contrast to other approaches to this problem, and in recognition of the limits of our responsibility attribution practices, we proposed to understand human-machine interactions of a sufficient degree of interactiveness and autonomy between the interactants as forming responsible collectives. In order to do so, we provided arguments that autonomous machines can meet at least one influential set of requirements for membership in collectives. We have shown that attributing any kind of responsibility to machines remains ontologically, normatively, and practically viable through collectives and their ability to bear responsibility. It also provides an explanatory perspective on the growing terminological variety within the debate surrounding normative questions of responsibility in human-machine interaction. Collective responsibility, as an available concept for explaining our normative practices, can be fruitfully deployed to relieve humans of being the sole bearers of responsibility in highly complex interactive contexts without making machines bearers of said responsibility.

Author Biographies

Niël H. Conradie (PhD, RWTH Aachen University) is a Postdoctoral Researcher in Applied Ethics at RWTH Aachen University. His research is focused on responsibility and trust and how these relate to AI and other emergent technologies, especially topics such as responsibility gaps and the possibility (and desirability) of trusting artificial systems.

https://orcid.org/0000-0002-1916-3171

Hendrik Kempt (PhD, RWTH Aachen University) is a Postdoctoral Researcher in Applied Ethics at RWTH Aachen University. His research is mainly on ethical implications of emerging human-machine interactions, foundational models, and digitized societies. He has published several books on these topics, including on the possibility of human-machine friendship (2022) and (un-)explainable technologies (2024).

https://orcid.org/0000-0002-5886-2987

Nils Freyer (MSc) is a research assistant at the Institute of Medical Informatics at the University Hospital RWTH Aachen. As part of his doctorate, he is working on the representation of uncertainty in clinical decision support systems. He has worked in the field of Natural Language Processing at the Aachen University of Applied Sciences and collaborated closely with the research associates of the Applied Ethics Group at RWTH Aachen University.

https://orcid.org/0000-0002-4460-3650

Saskia K. Nagel (PhD, University of Osnabrück) is full professor of Applied Ethics at RWTH Aachen University (Germany). Her research ranges from ethics of technology, ethics of "AI" and human-technology relations, to bioethics, the ethics of enhancement and philosophy of cognitive science. She works on questions of autonomy, responsibility, and trust in increasingly digitized societies, with a focus on human-technology relations.

https://orcid.org/0000-0001-9657-5121

References

Björnsson, G. (2017). *Explaining away epistemic skepticism about culpability* (Vol. 1). Oxford University Press. https://doi.org/10.1093/oso/9780198805601.003.0008

Björnsson, G. (2020). Group duties without decision-making procedures. *Journal of Social Ontology*, *6*(1), 127–139. https://doi.org/10.1515/jso-2020-0049

Björnsson, G., & Hess, K. (2017). Corporate crocodile tears? On the reactive attitudes of corporate agents. *Philosophy and Phenomenological Research*, *94*(2), 273–298. https://doi.org/10.1111/phpr.12260

Blomberg, O., & Hindriks, F. (2020). Collective responsibility and acting together. In S. Bazargan-Forward & D. Tollefsen (Eds.), *The Routledge handbook of collective responsibility* (1st ed., pp. 142–154). Routledge. https://doi.org/10.4324/9781315107608-12

Bradshaw, J. M., Feltovich, P. J., & Johnson, M. (2017). Human–Agent interaction. In G. A. Boy (Ed.), *The handbook of Human-Machine interaction* (1st ed., pp. 283–300). CRC Press. https://doi.org/10.1201/9781315557380-14

Braham, M., & van Hees, M. (2018). Voids or fragmentation: Moral responsibility for collective outcomes. *The Economic Journal*, *128*(612), F95–F113. https://doi.org/10.1111/ecoj.12507

Bryson, J. J. (2010). Robots should be slaves. In Y. Wilks (Ed.), *Natural language processing* (Vol. 8, pp. 63–74). John Benjamins Publishing Company. https://doi.org/10.1075/nlp.8.11bry

Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics*, *26*(4), 2051–2068. https://doi.org/10.1007/s11948-019-00146-8

Collins, S. (2019a). Collective responsibility gaps. *Journal of Business Ethics*, *154*(4), 943–954. https://doi.org/10.1007/s10551-018-3890-6

Collins, S. (2019b). *Group duties: Their existence and their implications for individuals*. Oxford University Press.

Conradie, N. H. (2023). Autonomous military systems: Collective responsibility and distributed burdens. *Ethics and Information Technology*, *25*(1), 20. https://doi.org/10.1007/s10676-023-09696-9

Corlett, J. A. (2001). Collective moral responsibility. *Journal of Social Philosophy*, *32*(4), 573–584. https://doi.org/10.1111/0047-2786.00115

Crootof, R., Kaminski, M. E., & Price II, W. N. (2022). *Humans in the loop* (SSRN Scholarly Paper No. 4066781). https://doi.org/10.2139/ssrn.4066781

Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology*, *18*(4), 299–309. https://doi.org/10.1007/s10676-016-9403-3

Dennett, D. C. (2014). When HAL kills, who's to blame? : Computer ethics. *When HAL Kills, Who's to Blame? : Computer Ethics*, 203–214. https://doi.org/10.1400/225034

Dung, L. (2023). Current cases of AI misalignment and their implications for future risks. *Synthese*, *202*(5), 138. https://doi.org/10.1007/s11229-023-04367-0

Edwards, A., Edwards, C., Spence, P. R., Harris, C., & Gambino, A. (2016). Robots in the classroom: Differences in students' perceptions of credibility and learning between "teacher as robot" and "robot as teacher." *Computers in Human Behavior*, *65*, 627–634. https://doi.org/10.1016/j.chb.2016.06.005

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge University Press.

Floridi, L. (2023). AI as agency without intelligence: On ChatGPT, large language models, and other generative models. *Philosophy & Technology*, *36*(1), 15. https://doi.org/10.1007/s13347-023-00621-y

Funk, M., & Coeckelbergh, M. (2020). (Technical) autonomy as concept in robot ethics. In J. L. Pons (Ed.), *Inclusive robotics for a better society* (pp. 59–65). Springer International Publishing. https://doi.org/10.1007/978-3-030-24074-5_12

Giubilini, A., & Levy, N. (2018). What in the world is collective responsibility? *Dialectica*, *72*(2), 191–217. https://doi.org/10.1111/1746-8361.12228

Gunkel, D. J. (2012). *Communication and artificial intelligence: Opportunities and challenges for the 21st century*. https://doi.org/10.7275/R5QJ7F7R

Gunkel, D. J. (2020a). *An introduction to communication and artificial intelligence*. Wiley.

Gunkel, D. J. (2020b). Mind the gap: Responsible robotics and the problem of responsibility. *Ethics and Information Technology*, *22*(4), 307–320. https://doi.org/10.1007/s10676-017-9428-2

Guzman, A. L., & Lewis, S. C. (2020). Artificial intelligence and communication: A Human-Machine Communication research agenda. *New Media & Society*, *22*(1), 70–86. https://doi.org/10.1177/1461444819858691

Hakli, R. A., & Mäkelä, P. A. (2016). Robots, autonomy, and responsibility. *What Social Robots Can and Should Do*, 145–154. https://doi.org/10.3233/978-1-61499-708-5-145

Hakli, R. A., & Mäkelä, P. A. (2019). Moral responsibility of robots and hybrid agents. *The Monist*, *102*(2), 259–275.

Hess, K. M. (2014). The free will of corporations (and other collectives). *Philosophical Studies*, *168*(1), 241–260. https://doi.org/10.1007/s11098-013-0128-4

Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology*, *11*(1), 19–29. https://doi.org/10.1007/s10676-008-9167-5

Himmelreich, J. (2019). Responsibility for killer robots. *Ethical Theory and Moral Practice*, *22*(3), 731–747. https://doi.org/10.1007/s10677-019-10007-9

Hindriks, F., & Veluwenkamp, H. (2023). The risks of autonomous machines: From responsibility gaps to control gaps. *Synthese*, *201*(1), 21. https://doi.org/10.1007/s11229-022-04001-5

Hoc, J.-M. (2000). From human-machine interaction to human-machine cooperation. *Ergonomics*, *43*(7), 833–843. https://doi.org/10.1080/001401300409044

Horstmann, A. C., & Krämer, N. C. (2020). Expectations vs. actual behavior of a social robot: An experimental investigation of the effects of a social robot's interaction skill level and its expected future role on people's evaluations. *PLOS ONE*, *15*(8), e0238133. https://doi.org/10.1371/journal.pone.0238133

Johnson, D. G. (2015). Technology with no human responsibility? *Journal of Business Ethics*, *127*(4), 707–715. https://doi.org/10.1007/s10551-014-2180-1

Kempt, H. (2020). *Chatbots and the domestication of AI: A relational approach*. Springer International Publishing. https://doi.org/10.1007/978-3-030-56290-8

Kempt, H., Heilinger, J.-C., & Nagel, S. K. (2022). "I'm afraid I can't let you do that, Doctor": Meaningful disagreements with AI in medical contexts. *AI and Society*. Scopus. https://doi.org/10.1007/s00146-022-01418-x

Köhler, S. (2020). Instrumental robots. *Science and Engineering Ethics*, *26*(6), 3121–3141. https://doi.org/10.1007/s11948-020-00259-5

Kompa, B., Snoek, J., & Beam, A. L. (2021). Second opinion needed: Communicating uncertainty in medical machine learning. *Npj Digital Medicine*, *4*(1), Article 1. https://doi.org/10.1038/s41746-020-00367-3

Königs, P. (2022). Artificial intelligence and responsibility gaps: What is the problem? *Ethics and Information Technology*, *24*(3), 36. https://doi.org/10.1007/s10676-022-09643-0

Levy, N. (2011). *Hard luck: How luck undermines free will and moral responsibility*. OUP Oxford.

Levy, N. (2018). Socializing responsibility. In K. Hutchinson, C. Mackenzie, & M. Oshana (Eds.), *Social dimensions of moral responsibility* (Vol. 1). Oxford University Press. https://doi.org/10.1093/oso/9780190609610.003.0008

List, C., & Pettit, P. (2011). *Group agency: The possibility, design, and status of corporate agents*. Oxford University Press.

Liu, H.-Y. (2016). Refining responsibility: Differentiating two types of responsibility issues raised by autonomous weapons systems. In C. Kreβ, H.-Y. Liu, N. Bhuta, R. Geiβ, & S. Beck (Eds.), *Autonomous weapons systems: Law, ethics, policy* (pp. 325–344). Cambridge University Press. https://doi.org/10.1017/CBO9781316597873.014

Ludwig, K. (2014). The ontology of collective action. In S. C. F. Hindriks & G. Preyer (Eds.), *From individual to collective intentionality: New essays*. Oxford University Press.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

McKenna, M. (2012). *Directed blame and conversation* (pp. 119–140) (D. J. Coates & N. A. Tognazzini, Eds.). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199860821.003.0007

Michalski, K., & Nowakowski, M. (2020). The use of unmanned vehicles for military logistic purposes. *Zeszyty Naukowe Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie. Ekonomika i Organizacja Logistyki*, *5*(4), 43–57.

Munch, L., Mainz, J., & Bjerring, J. C. (2023). The value of responsibility gaps in algorithmic decision-making. *Ethics and Information Technology*, *25*(1), 21. https://doi.org/10.1007/s10676-023-09699-6

Nyholm, S. (2018a). Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. *Science and Engineering Ethics*, *24*(4), 1201–1219. https://doi.org/10.1007/s11948-017-9943-x

Nyholm, S. (2018b). The ethics of crashes with self-driving cars: A roadmap, II. *Philosophy Compass*, *13*(7), e12506. https://doi.org/10.1111/phc3.12506

Nyholm, S. (2020). *Humans and robots: Ethics, agency, and anthropomorphism*. Bloomsbury Academic.

Oimann, A.-K. (2023). The responsibility gap and LAWS: A critical mapping of the debate. *Philosophy & Technology*, *36*(1), 3. https://doi.org/10.1007/s13347-022-00602-7

Oimann, A.-K., & Tollon, F. (2025). Responsibility gaps and technology: Old wine in new bottles? *Journal of Applied Philosophy*, *42*(1), 337–356. https://doi.org/10.1111/japp.12763

Pettit, P. (2001). *A theory of freedom: From the psychology to the politics of agency*. Oxford University Press.

Pettit, P. (2007). Responsibility incorporated. *Ethics*, *117*(2), 171–201. https://doi.org/10.1086/510695

Scanlon, T. M. (2015). Forms and conditions of responsibility. In R. Clarke, M. McKenna, & A. M. Smith (Eds.), *The nature of moral responsibility* (pp. 89–112). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199998074.003.0005

Shoemaker, D. (2015). *Responsibility from the margins*. Oxford University Press.

Smith, A. (2010). *The theory of moral sentiments*. Penguin.

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy*, *24*(1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Strawson, P. (1962). Freedom and resentment. *Proceedings of the British Academy*, *48*, 187–211.

Sundar, S. S., Jia, H., Waddell, T. F., & Huang, Y. (2015). Toward a theory of interactive media effects (TIME). In *The handbook of the psychology of communication technology* (pp. 47–86). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118426456.ch3

Sung, J., Grinter, R. E., & Christensen, H. I. (2010). Domestic robot ecology. *International Journal of Social Robotics*, *2*(4), 417–429. https://doi.org/10.1007/s12369-010-0065-8

Taylor, I. (2021). Who is responsible for killer robots? Autonomous weapons, group agency, and the military-industrial complex. *Journal of Applied Philosophy*, *38*(2), 320–334. https://doi.org/10.1111/japp.12469

Taylor, J., Yudkowsky, E., LaVictoire, P., & Critch, A. (2020). Alignment for advanced machine learning systems. In J. Taylor, E. Yudkowsky, P. LaVictoire, & A. Critch, *Ethics of artificial intelligence* (pp. 342–382). Oxford University Press. https://doi.org/10.1093/oso/9780190905033.003.0013

THeMIS Hybrid Unmanned Ground Vehicle. (2024). *Army technology*. https://www.army-technology.com/projects/themis-hybrid-unmanned-ground-vehicle/

Tigard, D. W. (2021). Responsible AI and moral responsibility: A common appreciation. *AI and Ethics*, *1*(2), 113–117. https://doi.org/10.1007/s43681-020-00009-0

Tollefsen, D. P. (2015). *Groups as agents*. John Wiley & Sons.

van de Poel, I., Nihlén Fahlquist, J., Doorn, N., Zwart, S., & Royakkers, L. (2011). The problem of many hands: Climate change as an example. *Science and Engineering Ethics*, *18*(1), 49–67. https://doi.org/10.1007/s11948-011-9276-0

Veluwenkamp, H. (2025). What responsibility gaps are and what they should be. *Ethics and Information Technology*, *27*(1), 14. https://doi.org/10.1007/s10676-025-09823-8

Watson, G. (2004). *Agency and answerability: Selected essays*. Clarendon Press.

Wringe, B. (2016). Collective obligations: Their existence, their explanatory power, and their supervenience on the obligations of individuals. *European Journal of Philosophy*, *24*(2), 472–497. https://doi.org/10.1111/ejop.12076

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019a). Algorithmic decision-making and the control problem. *Minds and Machines*, *29*(4), 555–578. https://doi.org/10.1007/s11023-019-09513-7

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019b). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology*, *32*(4), 661–683. https://doi.org/10.1007/s13347-018-0330-6

HUMAN-MACHINE
COMMUNICATION