Armando Di Stasio, Alessandro G. Buda, Giuseppe Primiero

Criteria for a theory of functional responsibility for AI as sociotechnical systems

(doi: 10.1422/120419)

Sistemi intelligenti (ISSN 1120-9550)
Fascicolo 1, aprile 2026

Ente di afferenza:
Università statale di Milano (unimi)

ARMANDO DI STASIO ALESSANDRO G. BUDA GIUSEPPE PRIMIERO

CRITERIA FOR A THEORY OF FUNCTIONAL RESPONSIBILITY FOR AI AS SOCIOTECHNICAL SYSTEMS

1. Introduction

While AI ethics is critical to the development of autonomous systems, incorporating ethical principles into them is controversial. Which specific moral theory and which values to implement is already a significant obstacle, as observed by Etzioni and Etzioni (2017). For the time being, the autonomy of AI systems is dubious at best, hence their outcomes are ultimately and at least partially attributable to humans. Accordingly, in the task of defining responsibility for AI as sociotechnical systems, it seems reasonable to work on the conditions and the environment in which these systems are developed, deployed, and used. Under this assumption, providing a causal account of the system's behavior by tracing the contributors that are responsible for its outcomes seems necessary.

The term "responsibility" may itself be too vague: according to Heinrichs (2022) it refers to the set of practices that allow one to answer the question:

In what way (1) is who (2) responsible for what (3) to whom (4) according to what norm (5) and for whose sake (6)?

The first and the second part of the question are typically answered by appealing to a dialogical assumption: when we attribute responsibility to someone we expect a justification, an explanation, or a reaction to being praised or blamed or, finally, an attempt to compensate the harm done. These concepts, in turn, often overlap in everyday talk, but they are clearly distinguishable, see Vincent (2010), Poel (2011) and van de Poel and Sand (2021):

- Accountability describes the reason and explanation-giving expectation that we have towards a person we hold responsible;
- Praise/blameworthiness indicates that we expect a person to react positively/negatively to our moral evaluation;
- Liability concerns the expectation of a person to compensate for the caused damage.

They are called *backward-looking* dimensions because they focus on events that have already happened and which we retrospectively investigate in order to evaluate who was actively involved and on what grounds the relevant actions have been carried out. *Forward-looking*, *i.e.*, prospective dimensions focus instead on the normative values that a person ought to hold in her future conduct, such as obligation and virtue. In the following we concentrate on the ascription of responsibility by tracing the contribution of the relevant agents involved.

We subscribe to the thesis that recipients for responsibility attribution should have "the ability to be attributed moral properties independently of external ascription" (Heinrichs, 2022, p. 6). However, we also claim that AI systems do not intrinsically have moral properties. Such a broad and apparently unqualified assertion seems to require a better justification for at least two reasons. First, the moral debate on AI is increasingly moving towards more specific considerations of the different kinds of systems, hence one should at least consider the specific moral questions raised by distinct contexts. For example, from the point of view of the moral implications, it seems that an LLM cannot be compared to a self-driving car (SDC) or to an autonomous weapon system (AWS), and not even to credit scoring or predictive policing systems. Second, even if denying the moral agency of AI systems, one should take into account that AIs can be considered to have a moral status in a passive sense, as subjects of moral concerns from moral agents (see, for instance, Sinnott-Armstrong, Conitzer, 2021).

However, in a similarly obvious fashion, certain facts about AI systems suggest that such qualification may be unnecessary. One should take into account that, despite their differences, LLMs, AWSs, predictive systems, and decision-support systems, all share the same goal and the same means: provide the most probable outcome/response given the data provided minimizing some error function. Moreover, although they do not raise the classic moral issues like SDCs or AWSs (e.g. trolley problem, responsibility gap, etc.), and despite their apparent harmlessness, LLMs have already been shown to pose a great risk to the mental health of users (see Moore *et al.*, 2025; Morrin *et al.*, 2025; Preda, 2025). This becomes particularly worrying when combined with statistical findings in which 20% of respondents declared that sentient AI systems currently exist (Anthis, Pauketat, Ladak, Manoli, 2025), or in which 17% of AI researchers and 18% of adults in the United States believe that at least one AI system has subjective experience (8% and 10%, respectively, believe that at least one AI system has self-awareness) (Dreksler *et al.*, 2025).

These and similar results show how much the discourse on artificial intelligence has been polluted by the so-called *scifitization*, the "conceptual drift" that leads "[r]eaders and viewers [to] have the impression that the term 'AI' refers to what they know from culture ('artificially intelligent' humanoids) rather than to what really exists in the world (smart hoovers or fallible chatbots)" (Porębski, Figura, 2025, p. 3).

In the context of AI ethics, scifitization takes the form of the *robotic conception* of AI, focusing on artefacts and their intrinsic properties, which inevitably leads to the definition of AI systems as artificial moral agents (AMAs), for which both active and passive moral status must be considered. However, we agree with Riesen (2025) that "[n]ot only is there little evidence that we need AMAs to have safe AI, but there are also strong moral reasons not to create them" (p. 8). Therefore, in this paper we endorse the *sociotechnical view* of AI technology, in which AI attributes derive from being designed or used in a certain manner, which are external ascriptions, in coordination with human agents. When an algorithm seems to show moral properties, it is actually to the concert of human design, usage, purpose, and perspective that is being attributed such morality.

More reasonably, an approach of attributing responsibility to AI consists of the establishment of a *system* responsibility for the outcomes of the actions executed by agents including humans and AIs. For example, Beck (2016) defines the legal person of robots as the "bundling of all the legal responsibilities of the different parties (users, sellers, producers, etc.)" (Heinrichs, 2022, p. 6). A similar approach is endorsed by Nyholm (2018), who reconceptualizes agency as collaborative or individual, supervised or deferential, domain-specific or domain-independent. Nyholm attributes collaborative, supervised, human-deferring agency to AI systems: this allows him to ascribe responsibility to the group of human collaborators employing the system. Furthermore, the inadequacy of the term "responsibility" once again emerges, since the task is now to assess "who is responsible for what aspects of the actions" (Nyholm, 2018, p. 1214). This, in turn, implies a differentiation of the types of responsibility and the reduction of its attribution to the human agents involved in the system, rather than simplistically bundling together several aspects which require specific evaluation, such as the role played by the designers, the budget constraints given by the financial team, or the deviant uses by the consumers.

Baum, Mantel, Schmidt, and Speith (2022) argue that the importance of the concept of responsibility and accountability lies in demanding an appropriate explanation of a particular conduct. In turn, this is desirable because it leads to considering an agent's reasoning and assessing possible misconceptions. More generally, keeping track of responsibility chains can be paramount in informational systems, which are prone to fall victim to biases and deviant uses. The appeal of this possibility lies in identifying which level or component of the system is to be held accountable for a certain outcome, not only for reprimanding purposes, but also for a more precise and efficient identification of misfunctions in the system.

On these lines, Primiero (2025) argues for a three level approach to responsibility attribution concerning the development and use of AI systems. The first level concerns who is responsible for creating a fair system: from the selection of training data, to the design of optimization functions and algorithms, as well as the training and validation phases. This level essentially belongs to the developer – understood as the complex category of all the actors involved in building a Machine Learning (ML) system. The second level concerns the responsibility of verifying the results of an ML system to ensure that the first-level responsibility has been met – that is, that the system has been trained and developed in a way that ensures fair results. Who is responsible for this verification phase mainly depends on political decisions regarding the regulation of AI development and its economic distribution. The third level concerns clients, users, and deployers, and how their choices impact the environment, the community, and the infrastructure.

Building on a similar viewpoint and drawing on Halpern (2016)'s distinction between type and actual causality, Dastani and Yazdanpanah (2022) introduce the concept of *actual responsibility*. This approach is intended to enable the tracing of AI systems' behavior and the attribution of responsibility to them, or to specific components which played a role in producing a given outcome. They contend that the agent-oriented perspective on responsibility can be viewed as an extension of the event-oriented approach, with events understood as actions determined by agents.

In the following, we will close up on a notion of *functional responsibility*, which considers events as *functions* executed by the *users* of AI systems, that allows (but does not oblige) for departure from ascription of morality to the agents involved, and which is focused on *tracing*, in the light of the notion of second-order responsibility mentioned above. To do this, we will first consider Matthias (2004)'s *responsibility gap* in sociotechnical systems and explain how and why it arises. We will then consider Floridi (2013, 2016)'s *faultless responsibility/mindless morality* which will serve as a working hypothesis, albeit with a slight modification in its approach and scope, providing a theoretical framework for discussing responsibility without necessarily implying intentionality or consciousness. To further refine this working hypothesis, we will incorporate de Sio and van den Hoven (2018)'s *meaningful human control*. In particular, we will focus on one of its two conditions – *traceability* – due to its non-reliance on assuming consciousness and its requirement for a precise account of the contributions that a group of agents has made to an outcome. We provide criteria for a such a notion of responsibility for sociotechnical systems with a similar characterization to Gebru *et al*. (2021) and Douer and Meyer (2023)'s quantitative approaches, by intertwining responsibility and information. The aim is to reconcile technical and ethical reasoning in the field of AI, with the aim of laying down the basis for a formalization. Formalization is relatively lacking in the current literature, especially when considering alternatives to the traditional approaches focusing on implementing ethical principles in artificial agents. Nonetheless, the operationalization of responsibility ascription and its verification will require methods to allow for implementation.

The rest of the paper is structured as follows. In Section 2 we illustrate various criteria for the notion of responsibility in particular with respect to its implications for morality ascription, and clarify our understanding of a functional notion of tracing for responsibility attribution. In Section 3 we specifically investigate a faultless notion of responsibility. In Section 4 we introduce four defining criteria for our notion of functional responsibility: amorality, distributability, quantifiability, surjectivity and compositionality. In Section 5 we sketch the next steps of this project in view of formal verification of responsibility ascription in the context of AI systems.

2. Conditions of Responsibility for AI as sociotechnical Systems

Understanding under which conditions responsibility can be attributed to AI systems is an essential step towards defining the possibility of any AI ethics. In turn, the landscape of various principles of responsibility determines which such conditions can or cannot be accepted.

Standard (but not uncontroversial) accounts link responsibility to the moral status of the responsible entity, focusing on its intrinsic properties (Liao, 2020), and leading to the need of a conception of AI systems as artificial moral agents, due to the trade-off between machine autonomy and human control. However, such a taken-for-granted trade-off, may be caused by misconceptions about autonomy and control, originated by the (bad) influence of popular culture in research programs, *i.e.*, the robotic conception of AI. As Riesen (2025) pointed out, philosophers typically view autonomy as a one-dimensional concept, capped by human-like moral autonomy, and infer that increasing autonomy necessarily reduces human control. In practice, however, this portrayal does not reflect reality, even in the most sophisticated AI systems, "because autonomy does not exist on a single dimension and increased *functional* autonomy need not imply decreased human control (Parasuraman, Sheridan, and Wickens (2000); Riesen (2022); Scharre (2018); Shneiderman (2022)" (p. 4). On the contrary, switching the attention from the research of the intrinsic moral feature of AI to the complex sociotechnical system[1] in which AI is embedded, together with humans, hardware, sensors *etc.*, provides both a more accurate description of AI systems and a safer research program.

[1] Riesen (2025) refers to *sociotechnological* systems in order to include only the individuals and organizations whose active participation is essential to achieving technological goals. We believe that this account can be extended to sociotechnical systems, including also policy makers, interest groups, marketers, advertisers *etc*.

In this context, the very aim of attributing responsibility to AI systems is subject to a structural problem. Matthias (2004)'s notion of *responsibility gap* illustrates how the black-box nature of ML systems poses an ethical, practical, and responsibility problem. How can we hold someone responsible – including the developers, designers and users – if the system's specification is opaque? This issue arises from the traditional accounts of moral responsibility attribution, which seem to be lacking with regards to artificial agents. The conditions that are most commonly assigned to human beings are the Control Condition (CC) and the Epistemic Condition (EC). They can constitute a benchmark for our analysis, as well as for any other.

The Control Condition states that an agent is responsible for an action or decision if and only if she has control over it (Fischer, Ravizza, 1991; Talbert, 2025). This condition aims to attribute moral responsibility to the agents who have a certain form of power, or control, over the investigated action. This is commonly regarded as the possibility of having done otherwise – as in Frankfurt (1969)'s Principle of Alternative Possibilities – which also seems to imply free will, that is, having chosen or decided to exercise such power over an action. This view characterizes most philosophical accounts of moral responsibility, dating back to Aristotle – as shown by Sakezles (2007) and Rowe (1987) – responsibility has been concerned with actions characterized by voluntariness or choice. To the extent that the control condition seems to entail conscious states, in this article we do not endorse such a concept of intentionality; however, we remain neutral to the possibility of AI having intentionality in a minimal sense of satisfying *aboutness* and *representation* of external contents, as defined in Jacob (2023). This choice does not mean however that outcomes produced by AI are not morally relevant, as Coeckelbergh (2020) shows. For example, the decision of a self-driving car to hit a pedestrian can be regarded as morally wrong, even though the causal factors are not.

The second condition that is normally ascribed to moral responsibility is the Epistemic Condition (EC), also called "knowledge, cognitive, or mental condition", as described by Rudy-Hiller (2022): for an agent to be praise/blameworthy for an action, she must be aware of certain things. This condition cannot be assumed to hold for AI systems because – again – "being aware" runs the risk of assuming conscious mental states, as Himma (2009) points out. Even with a formulation not utilizing the term "aware" – thus granting the satisfaction of the EC – traditional accounts of responsibility do not satisfy the first condition (CC). Since moral agency understood according to the traditional accounts implies consciousness as a necessary but not sufficient condition, and we refrain from ascribing the latter to artificial agents – see Johnson (2006) – we maintain a critical stance towards both conditions.

Other accounts for addressing responsibility in sociotechnical systems remain based on the conditions above, and their implications. But they significantly shift the focus of responsibility to those agents who can actually be attributed awareness or intentionality. A comprehensive approach to the responsibility gap which still endorses the control condition is Meaningful Human Control (MHC) (de Sio and van den Hoven, 2018; Hagendorff, 2020). MHC is the concept according to which humans, rather than AI systems, should ultimately be in control of, and therefore be morally responsible for, relevant decisions. According to Fischer and Ravizza (1991), a person is morally responsible for an action if and only if she has "guidance control" over that action. This in turn implies that the agent must act based on a decisional mechanism which is responsive to reasons (Reason Responsiveness), *i.e.*, such decisional system must be able to recognize the reasons for action (or omission) and bring about such action (or omission); and that the decisional mechanism must be an essential component of the agent's identity (Ownership Condition). This means that the agent must have *taken responsibility* for the process through which she makes decisions. de Sio and van den Hoven (2018) suggest that this account of human control over actions should also be extended to artifacts. Nozick (1981)'s analysis on the conditions upon which humans have knowledge is then called upon to introduce the concept of tracking, which can be defined as the reliable correspondence between an agent's mental states and the state of the world that ensures that humans can hold knowledge.

The authors point out that the system will therefore have to ensure a dual tracking process:

1. Human agents' beliefs must track the state of affairs of the world when the system is employed;

2. Relevant moral reasons of the relevant agents employing the mechanism must be kept track of and carried out.

This is to say that, not only in case the state of the world were different, but also in case the human agents' moral reasons were dfferent, the mechanism would track them and accommodate those variations. This leads the authors to formulate the first necessary condition for MHC:

Definition 1 (MHC_1). *A decision-making system must be responsive to the relevant human moral reasons of the relevant human agents. In short, it must track the relevant human moral reasons.*

The second condition for MHC is then based on the adaptation of Fischer and Ravizza (1991)'s ownership condition. This means that the humans collaborating with the AI systems must be put into the position to know what is actually happening within the artificial agent's mechanism, while also being aware of the moral impact of its usage. The process that enables the AI system's outputs to be traceable back to a human being who is appropriately and meaningfully aware of the system's capabilities and effects is called *tracing*:

Definition 2 (Tracing). *A human agent may be responsible for an outcome even if the conditions for responsibility are not satisfied at time* t_0 *on condition that she was responsible at* t_{-1}.

The main characteristic of the tracing condition is its applicability to sociotechnical systems where non-human agents are involved. While the standard formulation of the tracing condition based on temporality is a good approximation to determine the evolution of the actions which require responsibility attribution, it has a major shortcoming. Indeed, it does not determine explicitly whether the responsibility ascription can be morally loaded, in the presence of a self-aware agent or agent with moral intentionality. In order to address this aspect, instead of focusing on the temporal component, we shift the attention to the functions executed within the system by its defining agents. Functions can be defined as the role an agent (human or not) plays in bringing about an outcome. In turn, a functional notion of tracing can reconstruct the structural contributions that each agent can bring to the overall outcome of the system. The adjective "structural" serves the purpose of considering the hierarchical organization of the system, where each layer accomplishes a task while being connected to other agents' actions. Furthermore, temporal conditions may be difficult to pinpoint in the abstract, especially when the complexity of the system increases. Instead, the requirement of defining functions and their input/output relations, means indirectly to determine which agents provide input to which other agents. The reference to functions simplifies the process by paying attention to the implemented actions, thus guaranteeing a more direct approach. Another advantage consists of the possibility to indicate which levels or layers causally/functionally contributed to the outcome: not only does this enable a more thorough investigation of the causes of a harmful outcome by providing an overview of the information flow within the system, but it also aids in identifying which *kinds* of agents were involved and what roles they played by referring to the level they are located at. A function-based approach could first determine which level was responsible.

The second condition for MHC is defined by de Sio and Mecacci (2021) as follows:

Definition 3 (MHC_2). *The system must be traceable to a proper moral understanding on the part of one or more relevant human agents who design or interact with the system, meaning that there is at least one human agent in the design history or use context involved in designing, programming, operating and deploying the autonomous system who*

- *understands it or is in the position to understand the capabilities of the system*
- *understands it or is in the position to understand that others may have legitimate moral reactions toward them because of how the system affects the world.*

Hence, tracing does not directly avoid the issue of moral responsibility ascription. It rather imposes, in the form accepted by MHC, that such morality be identified in the correct agent, namely a human (Amoroso and Tamburrini, 2020). As shown in Amoroso, Garcia, and Tamburrini (2022), the urgency of establishing verifiable and agreed-upon human control requirements imposes considering humans as fail-safe actors contributing to the prevention of malfunctioning. Crucially, human control is also required to avoid accountability gaps: hence the role of humans as "accountability attractors". Indeed, by having a human in the loop, the conditions for attributing responsibility can – although not without difficulty – take place. Finally, the authors point out that human control also acts as a "moral agency enactor": life, death and human integrity decisions are not made by artificial agents. The authors suggest a *differentiated* and *prudential* approach, sensitive to the autonomy level of the agents and allowing exceptions to default, international rules, which are based on assigning low autonomy to all artificial agents. In our case, the functional tracing condition can be extended to require moral understanding by those agents who satisfy MHC_2, thus requiring that our analysis be a differentiated and prudential account too. For these reasons, we now propose a new definition of tracing based on de Sio and van den Hoven (2018)'s formulation by foregoing the temporal element and substituting it in terms of function as described above, one that allows for such functions to be occurring at different levels by different agents:

Definition 4 (Functional tracing). *An agent may be responsible for an outcome even if the conditions for responsibility are not satisfied at the current functional level of the outcome, but they are at a different level*.

This formulation can be linked to the need for "means [...] to link the AI systems' decisions to their input data and to the actions of stakeholders involved in the systems' decisions" (Dignum, 2020, p. 218), while also attempting to answer van der Loeff, Bassi, Kapila, and Gamper (2019)'s call for a more structural approach, rather than focusing solely on technical or legislative solutions.

In the following we further address the issue of sociotechnical systems where AI and human agents collaborate. The absence of intentionality and awareness in artificial agents has been described and explained above, along with the related problems of accountability and responsibility gaps. To close these gaps, it will be examined whether some kind of responsibility can be ascribed to human and artificial agents alike, while still preserving the intuition that, as designers and users of the systems, humans play a privileged role, and should therefore be appropriately held responsible.

3. Faultless responsibility

In distributed environments – systems consisting of collective actions in which moral responsibility can be allocated among multiple agents – the attribution of responsibility is ambiguous. A sociotechnical system comprising both human and artificial agents can cause Distributed Moral Actions (DMA): morally relevant actions which are caused by morally neutral interactions. As long as the exclusive focus is on intentionality, responsibility gaps emerge, and accordingly the problem of Distributed Moral Responsibility (DMR).

While not using the term "responsibility gap", Floridi (2016) stresses this issue as follows: focusing on individual intentionality leads to the attribution of individual responsibility. Individual intentionality is necessary for this standard approach because, otherwise, it would not be useful to punish those who commit crimes. However, individual intentionality on this account is closed neither under direct causal implication (if an agent *A* intends to cause *a* and *a* causes *b*, then *A* means to cause *b*), nor under distributed causal implication (if *A* intends to cause *a* and *B* intends to cause *b* and *a* and *b* jointly cause *C*, then *AB* intend to cause *C*), due to its monotonicity. Consequently, standard agent-oriented ethics faces a dilemma: either it ceases to concern itself with DMAs and DMRs, or it must reduce both of them to individual, intentional moral actions.

Floridi considers this as unsatisfactory, invoking a patient-oriented ethics, focused on whether an action influences a relevant group of individuals. Shifting the attention to the receivers of the executed actions – that is, to the victims – has the advantage of not only dropping intentionality as a necessary condition, but it also takes into account how the relevant components of a sociotechnical system should be influenced in order to facilitate right actions and stifle wrong ones.

Floridi suggests that "an ethics of state transitions, independent of the intentions of the agents involved, can provide a full account of DMR" (Floridi, 2016, p. 6). The only assumption for such an ethical account would be that some *states of the system* are morally better than others, and therefore to be pursued *for the sake of the system itself*. This holistic view, which seems in line with the differentiated and prudential account advocated by Amoroso and Tamburrini (2020), is reasonably apt to describe a sociotechnical system where AI may be responsible without needing intentionality as a defining condition.

Two things are worth noting. First, it is not necessary to identify the intentions behind the agent's actions. Second, this kind of strict liability, which Floridi calls *faultless responsibility*, is applied to the individual agents, albeit distributed to many, and not to the whole system. Faultless responsibility, then, means that agents are responsible without necessarily being at fault: that is to say that agents are held responsible because they are *causally* involved in bringing about an undesirable system state, and not because they intended to do so. One could go as far as saying that each agent is responsible *by default* for what the sociotechnical system achieves. This is relevant for *taking responsibility*: in the absence of exculpation mechanisms, the developers and users of the system will have to take a forward-looking responsibility in order to promote values and norms so that no harmful system states are achieved.

This notion of faultless responsibility can be put into question by Redaelli (2024)'s notion of preterintentionality. While Redaelli focuses on generative AI, it can be valuable to compare his account of intentionality with Floridi's, in order to gain a more defined working hypothesis. Redaelli challenges the designer fallacy, which consists of the reduction of all functions of a certain technology to the designer's intentions, as in Johnson (2006), arguing that generative AI goes *beyond* human intentions (users and designers alike): AI systems do not merely reproduce the provided inputs, but actually have a margin of indeterminacy that allows for unexpected outcomes. This *surplus* is nonetheless desired by the designers, who often resort to ML in order to ensure original and unseen outputs (which may be in the form of text or any other digital medium). Redaelli refers to this using the term "preterintentionality", which highlights the irreducibility of AI intentionality to human intentionality. This term captures the surpassing of human intention while nonetheless being dependent on it: it *emerges* from the conspicuous interaction between human and AI.

Floridi and Redaelli's proposals share some fundamental similarities. Firstly, they agree on the fact that outcomes resulting from AI often lack intentionality as a cause: while Floridi focuses on the legal and ethical aspect of this absence, Redaelli highlights what we could regard as an epistemic gap, where the outcomes of artificial agents are not only unknown, but the reality they are embedded in is constructed by them.

Secondly, both accounts agree on the shortcomings of intention-based approaches, and therefore suggest two ways in which responsibility can be attributed without strict moral agency, *i.e.*, faultless responsibility and preterintentionality. The former accounts for the attribution of responsibility despite lack of negligence or (malicious) intent, while the latter ascribes intentionality as an emergent feature of the *interaction* between humans and artificial agents. Finally, intentionality is neither a sufficient, nor a necessary condition for responsibility according to both: while Redaelli leaves open the question of what kind of responsibility one could attribute to artificial agents, Floridi proposes a forward-looking account of responsibility, which aims at preventing harm by holding all agents involved in sociotechnical systems responsible by default, even when no clear fault exists. Redaelli's notion of preterintentionality does not dissolve responsibility, quite the contrary: it demands more adequate responsibility accounts. Furthermore, Redaelli does not exclude a tracing account of *causal contributions* – and therefore causal responsibility – to an outcome. The current work focuses on the latter, and not the former: we are primarily interested in functions, rather than intentions.

One can agree on the untraceability of artificial agents' functions to human *intentionality* while still attempting to trace the relevant, purely causal links involved in an outcome. Furthermore, if one were to consider human intentionality from this point of view, it would follow that it would turn out to be irrelevant, following Floridi: responsibility persists despite lack of fault and, therefore, intention. An objection might arise: it could be noted that a hidden assumption is at play in the current project, namely that all causal links are traceable. It is clear that black box AI systems are opaque and therefore offer little to no possibility of tracing. Nonetheless, from a systemic and backward-looking point of view, the fact that non-transparent methods have been implemented into an AI system can be traced back to a design choice. From this perspective, the term preterintentionality seems to offer an incomplete description of the phenomenon, as it is related only to actions whose results go beyond the agent's intentions. Roman law offers a companion term to describe the behavior of an AI system designer: *dolus eventualis*. This term indicates that the results of an action, even if unintended, are accepted as possible consequences: *i.e.*, the agent, even if not directly intending them, accepts them as a risk. Of course, whether or not the human designer(s) knew about the risk of inexplainable methods employed is relevant to moral responsibility, but not to the attribution of strict causal responsibility *within the system*.

The fact that artificial agents are not held responsible is crucial to our inquiry on the responsibilities within a sociotechnical system. Indeed, the distinction between causal accountability and moral responsibility is fundamental to trace which (groups of) agents, and to what extent, are responsible. A minimal sense of agency can be captured by the concept of function. More specifically, we refer to *functions* as the focus of the tracing account. With *function* we identify the contribution that *any* agent provides to the overall operations that the system puts into place. From a purely descriptive perspective, a function can be expressed as a simple input/output relation that the agent facilitates. If a causal account is required, such description should be enhanced with a mechanistic interpretation to explain which components permit the I/O relation to occur and in which order of execution and composition. By relying on such functional reading, we allow an approach which is not concerned with intentionality of any kind, rather simply with the contributions that an agent – be it human or artificial – makes to an outcome. In this sense, the term function is preferable due to its technical nature: certain features of a system are the result of the ramification of several agents which play a role in bringing them about. The task (or action or I/O relation in the sense described above) that an agent carries out is thus its function. The neutrality of such a term allows for a more relaxed notion of agency while still trying to account for traceability.

We will use the term *functional responsibility* to denote the fact that an agent (moral or not) has (causally) contributed to a system's function, without implying intentionality. *Moral responsibility*, by contrast, will refer to a moral, and therefore human agent who can be morally blamed for the implemented function because of the inherent intentionality of her actions. In this sense we will refrain from addressing the DMA problem, as it may concern only a part of the system, namely that which includes agents equipped with intentionality. A neutral term which will be used and justified in the final section below to account for *any* kind of agent (moral or amoral, artificial or natural, active or patient) is *user*. This choice counters the widespread adoption of the term *agent* in the current AI terminology, especially in view of the current *Agentic AI* trend. While unusual, we stick to it in order to be able to separate agency as the exercise or manifestation of a capacity which in the philosophy of action is construed in terms of intentionality, and the latter is explained in terms of causation by the agent's mental states (Schlosser, 2019). As functional responsibility wishes to refrain from accounting for the intentionality of actions while admitting that some of its component may be intentionally loaded (and in this it differs from standard technical artefacts), our choice of the term *user* is meant to include agents, but not be limited to them, nor to be constrained to passive forces only.

4. Four essential Criteria for Functional Responsibility

In this section we formulate four basic desiderata at the basis of our notion of functional responsibility for a hybrid sociotechnical system.

4.1. *Amorality*

Functional responsibility does not assume intentionality with respect to *artificial* agents. The ascription of moral agency – which, in turn, would imply intentionality and consciousness – to artificial agents would allow humans to discharge their responsibilities to AI systems, exculpating themselves. This also avoids the bold claim of assigning consciousness to artificial agents. Thus, among our desiderata, amorality – the lack of moral agency – occupies a prime position.

To further illustrate this point, let us consider de Sio and van den Hoven (2018)'s notion of tracking as opposed to tracing. Tracking is based on *responsiveness* to moral reasons, grounded on the *knowledge* of some states of the world and the related relevant moral rules. On the one hand, knowledge implies a human, and possibly moral agent involved, excluding *a priori* artificial agents; instead, we aim to develop a comprehensive account, based on the more primitive and impersonal notion of *information*: being functionally responsible is not an epistemic condition. On the other, being responsive to moral reasons may be seen as requiring moral reasoning, even if simply to identify what qualifies as such, and this would lead to the ascription of intentionality. Having knowledge of, and being responsive to the states of affairs *and* moral reasons runs the risk of appealing to intentionality.

Accordingly, our notion of functional responsibility satisfies the functional tracing condition: a neutral, algorithmically-computable, information-based concept focused on the *causal* contribution of an agent to an outcome, without any underlying assumption of intention. Intentionality may be used to explain human choices, for example the designer's decision to use a biased dataset, but it must not be required if the intent is to define responsibility for hybrid systems. Floridi's notion of faultless responsibility is thus accepted, but with a modification: instead of a patient-oriented approach, we adopt a user-oriented approach. The set of users includes not only agents and patients, but also all those informational sources and means of transmission, interpretation, and manipulation of information, available for agents' actions and/or allowing patients' conditions. In other words, by user we mean an "informational entity not necessarily able to *act on*, but still able to *use* (and *be used* as a source of) information" (Buda, Primiero, 2025, p. 10), hence including data artefacts and computational systems besides humans.

Not only does this prevent us from focusing on the agent/patient dialectics – which could lead us to consider their intentions – but it also lets us shift the attention to the purely *causal* contributions of the users. The term *function* is thus employed to underline the amorality of this concept of responsibility. *Function* refers to the causal contribution of a user – be it artificial or human to the sociotechnical system's outcome. It is the task or process that such user carries out, the role that it plays in bringing about the end result. On the other hand, to preserve tracking for morally responsible (human) agents, a distribution of responsibility is required, among users of different kind.

4.2. *Distributability*

While Floridi's notion of *faultless responsibility* holds *all* agents accountable by default, an objection could be raised: it seems unfair. Floridi anticipates this objection and solves it by appealing to already established legal practices (strict liability) and forward-looking responsibility: by considering every member responsible by default, all agents will be more prone to take responsibility and anticipate undesirable outcomes. We claim that what is needed is a qualification of the distribution of responsibility among the components of a sociotechnical system.

To this aim, a theory of the ontology and epistemology of computational systems (including hybrid sociotechnical systems) which allows one to distinguish among different types of users and how they interact in their functional operations is the User Levels (ULs) theory introduced in Buda and Primiero (2024). Traditional views of computational artefacts – centered on Levels of Abstraction (LoAs) – overlook essential pragmatic and contextual dimensions, especially in modern computing phenomena like deviant uses, esoteric languages, software development practices, and data-driven systems. ULs are introduced as a pragmatic complement to LoAs. Unlike the hierarchical, specification-centered LoAs, ULs capture how users impose their own semantic and normative expectations, which may even conflict with logical specifications. ULs form an evolving hierarchy of interpretation and use, reflecting varied user roles across different contexts. A pragmatic notion of computational act – integrating how an artefact is used, interpreted, and valued in context – is the basis to redefine *implementation* and *correctness* pragmatically: not solely in formal or structural terms, but also in relation to user-dependent semantics and norms. The theory addresses cases where behavior depends on context or usage, not just on formal specification. Computational artefacts are akin to semiotic systems shaped by practitioners, users, and contexts, not just formal sign structures. In this light, computing systems cannot be fully understood without recognizing how users interpret, value, and sometimes subvert specifications in real-world contexts.

UL theory allows one to account for Distributed Functional Responsibility (DFR) by considering different components of sociotechnical systems and their functional role, relevance, requirements, and norms where morality is included in the case of human agents. In this sense, each UL becomes responsible for a well-specified and traceable subset of actions. Each such computational act will be identifiable as idiosyncratic, innovative, act of expert redesign or of product design. Accordingly, responsibility for computational acts falling under the product design category will be traceable to the highest user level and shared below to all user levels involved by the system; expert redesign will be traceable to forms of responsibility shared across user levels; innovative computational acts will be traced to responsibility shared among users within a user level; idiosyncratic acts to individual users within a user level. For each such qualification, morality can be either included or excluded, depending on whether the relevant UL includes or not human agents.

4.3. *Quantifiability*

The qualification of functional responsibility dictated by ULs aids in making the attribution of responsibility fairer – by transparently showing the quality and the degree to which users have contributed. Moreover, it makes it compatible with an algorithmic approach to the tracing condition of MHC.

Tracing requires human agents (for which morality is involved) to be linkable to the system's outcomes: given a formal representation of the system's structure – including users, informational states and relations – it must be possible to determine algorithmically whether a user u or a group of users U within a user level, or an entire UL, or a set of distinct ULs is responsible for a formula ϕ. This distributed accountability in turn will ensure verifiability and praise/blameworthiness. It makes it possible to compute a user's degree of responsibility for the system's outcome. UL Theory has been formalized as a weighted modal logic in Buda and Primiero (2025) and it allows for an in-depth analysis through the reference to specific user levels. The underlying Logic of Pragmatic Information (LPI) provides a formalization of a whole sociotechnical system, rather than focusing on the moral judgments implemented. Other actors along with AIs come into play: designers, developers, owners, end-users. LPI lends itself well to our aim of providing a precise tracing of responsibilities within the system. This is because it lays out in a clear and transparent manner the hierarchies of the system, the users that act within them, and the information flow, captured by abstraction, instruction and implementation relations, thus making it possible to take tracing into account.

4.4. *Surjectivity and Compositionality*

Distributability guarantees that responsibility can be shared among more than one user, and quantifiability that each user responsible receives the right amount of praise or blame. A stronger condition would guarantee that each outcome of the system has at least one user held responsible for it, and that each partial contribution to an output is traced along different ULs.

The last of our desiderata is thus that responsibility be defined as a surjective and compositional function of users, with values in the set of possible outcomes of the system. These formal properties guarantee that functional responsibility can be represented as a computable function such that every outcome of the system is traced back to at least one user responsible for it, possibly identifying any chain of responsibility that led to that outcome. This requires us to formalize the relations among all the components of an AI system, spanning from designers to final users, across training and test samples, training engine, learned model, model output etc.

The ontology introduced in Buda and Primiero (2024) and formalized in Buda and Primiero (2025) provides a semantics in which accessibility relations of abstraction and implementation hold between hierarchically ordered user levels. These relations are weighted as a function of the role and influence that each user, and each level contributes to the final output. Based on these relations, one can trace back from the output to the source each component involved, ensuring that there are no responsibility gaps in the formalization of the system.

5. Conclusions

In this paper we presented a characterization of functional responsibility for hybrid sociotechnical systems, on the basis of faultless responsibility, traceability for meaningful human control, an ontology based on user levels and a logic of pragmatic information. It is characterized by not requiring morality, as being distributable across human and artificial agents, whose factual contribution can be quantified, and as guaranteeing that the system's set of possible outcomes is associated with one or more users across different levels. In addition to solving the problem of responsibility gap and many hands, it also bridges event-based responsibility and agent-based responsibility, resulting in a weak form of *actual responsibility* (Dastani and Yazdanpanah, 2022), deprived of (but not incompatible with) any epistemic, motivational, normative, and moral considerations, and providing a sufficiently flexible and granular tool to formalize highly contextual phenomena such as the behavior of AI systems. The extension of LPI with a Functional Responsibility operator which quantifies over sets of users and their relations is left for further work.

Acknowledgments

This research was supported by the Ministero dell'Università e della Ricerca (MUR) through PRIN 2022 Project SMARTEST – Simulation of Probabilistic Systems for the Age of the Digital Twin (20223E8Y4X), and through the Project "Departments of Excellence 2023-2027" awarded to the Department of Philosophy "Piero Martinetti" of the University of Milan.

References

Amoroso, D., Garcia, D., Tamburrini, G. (2022, 01). The weapon that mistook a school bus for an ostrich: Vulnerabilities in artificial intelligence and challenges for the meaningful human control of autonomous weapons. *Science & Diplomacy*, doi: 10.1126/scidip.ade6750.

Amoroso, D., Tamburrini, G. (2020, December). Autonomous weapons systems and meaningful human control: Ethical and legal issues. *Current Robotics Reports*, *1*, 4, pp. 187-194, https://doi.org/ 10.1007/s43154-020-00024-3.

Anthis, J.R., Pauketat, J.V., Ladak, A., Manoli, A. (2025). Perceptions of sentient ai and other digital minds: Evidence from the ai, morality, and sentience (aims) survey. In *Proceedings of the 2025 chi conference on human factors in computing systems*. New York, NY, USA: Association for Computing Machinery, https://doi.org/10.1145/ 3706598.3713329.

Baum, K., Mantel, S., Schmidt, E., Speith, T. (2022, February). From responsibility to reason-giving explainable artificial intelligence. *Philosophy & Technology*, *35*, 1, pp. 12, https://doi.org/10.1007/ s13347-022-00510-w.

Beck, S. (2016). The problem of ascribing legal responsibility in the case of robotics. *AI & society*, *31*, 4, pp. 473-481. Retrieved from https:// doi.org/10.1007/s00146-015-0624-5.

Buda, A.G., Primiero, G. (2024). A pragmatic theory of computational artefacts. *Minds & Machines*, *34*, pp. 139-170.

Buda, A.G., Primiero, G. (2025). A logic for using information. *Logique et Analyse*, *265*.

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

Dastani, M., Yazdanpanah, V. (2022, June). Responsibility of ai systems. *AI Soc.*, *38*, 2, pp. 843-852, https://doi.org/10.1007/s00146-022-01481-4.

de Sio, F.S., Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology*, *34*, 4, pp. 1057-1084, https://doi.org/10.1007/ s13347-021-00450-x.

de Sio, F.S., van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, *5*, 15.

Dignum, V. (2020). Responsibility and artificial intelligence. In M.D. Dubber, F. Pasquale, S. Das (Eds.), *The Oxford handbook of ethics of ai*. Oxford: Oxford Academic, pp. 215-230, https://doi.org/10.1093/oxfordhb/9780190067397.013.12.

Douer, N., Meyer, J. (2023). *Quantifying retrospective human responsibility in intelligent systems*, https://arxiv.org/abs/ 2308.01752.

Dreksler, N., Caviola, L., Chalmers, D., Allen, C., Rand, A., Lewis, J., ... Sebo, J. (2025). *Subjective experience in ai systems: What do ai researchers and the public believe?*, https://arxiv.org/abs/2506.11945.

Etzioni, A., Etzioni, O. (2017). Incorporating ethics into artificial intelligence. *The Journal of Ethics*, *21*, pp. 403-418.

Fischer, J.M., Ravizza, M. (1991). Responsibility and inevitability. *Ethics*, 2, pp. 258-278.

Floridi, L. (2013). The morality of artificial agents. In *The ethics of informa-*

tion. Oxford: Oxford University Press, https://doi.org/10.1093/acprof:oso/9780199641321.003.0007.
Floridi, L. (2016). Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*, 2083, pp. 20160112, https://royalsocietypublishing .org/doi/abs/10.1098/rsta.2016.0112.
Frankfurt, H.G. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy*, *66*, 23, pp. 829-839, https://doi.org/10.2307/2023833.
Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., III, H.D., Crawford, K. (2021). *Datasheets for datasets*, https://arxiv.org/abs/1803.09010.
Hagendorff, T. (2020, March). The ethics of ai ethics: An evaluation of guidelines. *Minds and Machines*, *30*,1, pp. 99-120, https://doi.org/10.1007/s11023-020-09517-8.
Halpern, J.Y. (2016). *Actual causality*. Cambridge: MiT Press.
Heinrichs, J.-H. (2022). Responsibility assignment won't solve the moral issues of artificial intelligence. *AI and Ethics*, *2*, 4, pp. 727-736, https://doi.org/10.1007/s43681-022-00133-z.
Himma, K.E. (2009, March). 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, pp. 19-29, https://doi.org/10.1007/s10676-008-9167-5.
Jacob, P. (2023). Intentionality. In E.N. Zalta, U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Spring 2023 ed.). Metaphysics Research Lab, Stanford University.
Johnson, D.G. (2006, November). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology*, *8*, 4, pp. 195-204, https://doi.org/10.1007/s10676-006-9111-5.
Liao, S.M. (2020, 09). The moral status and rights of artificial intelligence. In *Ethics of artificial intelligence*. Oxford: Oxford University Press, https://doi.org/10.1093/oso/9780190905033.003.0018.
Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology*, *6*, pp. 175-183.
Moore, J., Grabb, D., Agnew, W., Klyman, K., Chancellor, S., Ong, D.C., Haber, N. (2025, June). Expressing stigma and inappropriate responses prevents llms from safely replacing mental health providers. In *Proceedings of the 2025 acm conference on fairness, accountability, and transparency*. ACM pp. 599-627, http://dx.doi.org/ 10.1145/3715275.3732039.
Morrin, H., Nicholls, L., Levin, M., Yiend, J., Iyengar, U., DelGuidice, F., ... Pollak, T. (2025). *Delusions by design? How everyday ais might be fuelling psychosis (and what can be done about it)*. Retrieved 2025-08-22, from osf.io/preprints/psyarxiv/cmy7n_v6.
Nozick, R. (1981). *Philosophical explanations*. Cambridge: Cambridge University Press.
Nyholm, S. (2018). Attributing agency to automated systems: Reflections on human-robot collaborations and responsibility-loci. *Science and Engineering Ethics*, *24*, 4, pp. 1201-1219, https://doi.org/10.1007/s11948-017-9943-x.

Parasuraman, R., Sheridan, T., Wickens, C. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, *30*, 3, pp. 286-297.
Poel, I. (2011, 08). The relation between forward-looking and backward-looking responsibility, https://doi.org/10.1007/978-94-007-1878-4_3.
Porębski, A., Figura, J. (2025, Oct 28). There is no such thing as conscious artificial intelligence. *Humanities and Social Sciences Communications*, *12*, 1, pp. 1647, https://doi.org/10.1057/s41599-025-05868-8.
Preda, A. (2025). Special report: Ai-induced psychosis: A new frontier in mental health. *Psychiatric News*, *60,* 10, null, https://doi.org/ 10.1176/appi.pn.2025.10.10.5.
Primiero, G. (2025). Justice at the Core of a New AI Paradigm. *Harvard Data Science Review*, *7*, 4, https://doi.org/10.1162/99608f92.3f78ed8e.
Redaelli, R. (2024, July). Intentionality gap and preterintentionality in generative artificial intelligence. *AI & Society*, https://doi.org/10.1007/s00146-024-02007-w.
Riesen, E. (2022). The moral case for the development and use of autonomous weapon systems. *Journal of Military Ethics*, *21*, 2, pp. 132-150. Retrieved from https://doi.org/10.1080/15027570.2022.2124022.
Riesen, E. (2025, Apr 28). A sociotechnological-system approach to ai ethics. *AI & Society*, https://doi.org/10.1007/s00146-025-02335-5.
Rowe, W. (1987). Two concepts of freedom. *Proceedings and Addresses of the American Philosophical Association*, *61*, September, pp. 43-64.
Rudy-Hiller, F. (2022). The Epistemic Condition for Moral Responsibility. In E.N. Zalta, U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Winter 2022 ed.). Metaphysics Research Lab, Stanford University, https://plato.stanford.edu/archives/win2022/entries/moral-responsibility-epistemic/.
Sakezles, P.K. (2007). Aristotle and chrysippus on the psychology of human action: Criteria for responsibility. *British journal for the history of philosophy*, *15*, 2, pp. 225-252.
Scharre, P. (2018). *Army of none: Autonomous weapons and the future of war*. New York: WW Norton & Company.
Schlosser, M. (2019). Agency. In E.N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2019 ed.). Metaphysics Research Lab, Stanford University, https://plato.stanford.edu/archives/ win2019/entries/agency/.
Shneiderman, B. (2022). *Human-centered ai*. Oxford: Oxford University Press, https://doi.org/10.1093/oso/9780192845290.001.0001.
Sinnott-Armstrong, W., Conitzer, V. (2021, 08). How much moral status could artificial intelligence ever achieve? In *Rethinking moral status*. Oxford: Oxford University Press, https://doi.org/10.1093/oso/9780192894076.003.0016.
Talbert, M. (2025). Moral Responsibility. In E.N. Zalta, U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Fall 2025 ed.). Meta-physics Research Lab, Stanford University, https://plato.stanford.edu/archives/fall2025/entries/moral-responsibility/.
van de Poel, I., Sand, M. (2021). Varieties of responsibility: two problems of responsible innovation. *Synthese*, *198*, 19, pp. 4769-4787, https://doi.org/10.1007/s11229-018-01951-7.
van der Loeff, A.S., Bassi, I., Kapila, S., Gamper, J. (2019). AI ethics for sys-

temic issues: A structural approach. *CoRR*, *abs/1911.03216*, http://arxiv.org/abs/1911.03216.
Vincent, N. (2010). A structured taxonomy of responsibility concepts. In *Moral responsibility: Beyond free will and determinism*, https://doi.org/10.1007/978-94-007-1878-4_2.

Criteria for a theory of functional responsibility for AI as sociotechnical systems

Establishing criteria for assigning responsibility within sociotechnical systems involving AI technology is currently one of the fundamental tasks and problems of critical and formal reflection on technology. This article proposes a definition of functional responsibility that combines tracing and faultless responsibility in an ontology of sociotechnical systems structured on user levels. On this basis, four fundamental criteria are defined. First, responsibility must be assignable to artificial agents without moral implications, and only conscious users must be considered morally responsible. Second, responsibility for a system functionality must be distributable across multiple users. Third, it must be possible to quantify each user's contribution to each function of interest. Fourth, the responsibility function is surjective, meaning there are no outputs of actions without responsible users, and compositional, meaning the output of one responsibility function can be the input of another responsibility function. This work is preliminary to the formulation of a formal system for automatic responsibility verification.
Keywords: Functional responsibility, Sociotechnical systems, Philosophy of AI.

Armando Di Stasio, Department of Philosophy "Piero Martinetti", University of Milan, Via Festa del Perdono 7, 20122 Milano
distasio@studenti.unimi.it
https://orcid.org/0009-0001-8635-2327

Alessandro G. Buda, Department of Philosophy "Piero Martinetti", University of Milan, Via Festa del Perdono 7, 20122 Milano
alessandro.buda@unimi.it
https://orcid.org/0000-0002-5547-7514

Giuseppe Primiero, Department of Philosophy "Piero Martinetti", University of Milan, Via Festa del Perdono 7, 20122 Milano
giuseppe.primiero@unimi.it
https://orcid.org/0000-0003-3264-7100