

Abstract

From Minimal Procedural Addressability to Accountability: Conditions for Attributing Responsibility to Artificial Intelligence Systems

The subject of the dissertation is an analysis of the conditions under which one may consider attributing weak procedural responsibility and strong accountability to artificial intelligence systems. The main research question concerns which recognizable conditions of minimal addressability an artificial intelligence system should meet for weak procedural responsibility to be attributed to it, and which additional conditions of stabilization, strong accountability, and institutional application determine the scope of that responsibility. The dissertation distinguishes the responsibility of artificial intelligence systems from both full human moral responsibility and the neutrality of a technical tool, arguing that a threshold model is more appropriate. Minimal procedural addressability is the first condition for formulating claims toward a system for justification, assessment, and correction; however, it does not determine full moral subjecthood, legal personhood, or independent institutional accountability.

The dissertation begins from the diagnosis that contemporary artificial intelligence systems, especially adaptive and generative systems, reveal the limitations of classical models of responsibility. Their actions may have social, legal, and moral significance, although the processes leading to decisions are often opaque, distributed, and difficult to reconstruct. A tension therefore arises between the impact of systems on goods protected by norms and the traditional requirement that responsibility be attributed to a subject capable of intentional action, understanding norms, and being subject to justified criticism. The dissertation does not assume, however, that the mere social impact of artificial intelligence systems is sufficient to recognize them as moral subjects. The decisive question is whether the system's decisions and justifications can be incorporated into public practices of asking "why?", assessing reasons, and correcting actions. In this sense, "procedural addressability" means the minimal possibility of treating the system as a point of reference in procedures of justification, audit, and decision correction.

Methodologically, the work is analytical and conceptual. The author conducts no empirical research or experiments and does not design technical solutions. Technological examples serve

illustrative and heuristic purposes. In the case of IDx-DR/LumineticsCore, published clinical findings and regulatory documentation provide evidence only for descriptive claims about the system; applying H1–H3 + WI remains a conceptual analysis and does not empirically validate the model. The basic methods are conceptual analysis, argument reconstruction, comparative assessment of positions, and testing the model's coherence. The work reconstructs disputes about subjecthood, free will, consciousness, agency, and responsibility, and then applies their findings to the philosophy of artificial intelligence. It draws on literature in moral philosophy, action theory, philosophy of mind, ethics of artificial intelligence, philosophy of technology, and artificial intelligence regulation. References to legal documents serve to verify the model's institutional applicability; they do not, however, replace philosophical argumentation.

Chapter 1 defines the research problem, working hypotheses, and scope of the work. The dispute centers on whether artificial intelligence systems are merely tools or, under certain conditions, may function as quasi-subjects in practices of responsibility. The chapter clearly distinguishes between causal and normative agency. Responsibility cannot be attributed solely based on the effects of an action. Three main hypotheses and an institutional condition are introduced. H1 defines minimal procedural addressability as the basis of weak procedural responsibility. H2 treats functional affectivity as a possible mechanism for stabilizing choices consistent with reasons, while K4 separately assesses diachronic stability and the reliability of the system's behavioral profile. H3 defines the threshold of strong accountability, which depends on metacognitive control and explicit resolution of conflicts between reasons. The WI condition sets the limits for applying the H1–H3 model in public, organizational, and regulatory practices. The model is prospective: it specifies conditions for possible responsibility rather than establishing that current artificial intelligence systems satisfy them.

Chapter 2 reconstructs the concept of subjecthood in selected philosophical accounts. The analysis covers three functions of the subject, namely being a bearer of states, being an agent of actions dependent on reasons, and being an addressee of claims and responsibility. The minimal conditions of practical responsibility do not require a strong ontology of the “deep self.” What is crucial is the public recognizability of the addressee, the possibility of justifying action, and reason-responsive control. The second-person perspective is especially important. Responsibility means participation in practices of mutually demanding reasons, responses, criticism, justification,

and correction. The chapter also considers the distinction between phenomenal consciousness and access consciousness. Phenomenal consciousness is important in cases of experiential harm, but it is not a necessary condition for minimal procedural addressability. H1 requires the functional recognition of norms, anticipation of the consequences of violating them, counterfactual responsiveness to reasons, and the provision of sufficient and reliable grounds for decisions that allow their connection to the outcome to be assessed and enable audit, challenge, and correction.

Chapter 3 defines diagnostic and operational criteria of artificial moral subjecthood. Criteria K1–K3 support the assessment of whether H1 is satisfied, whereas K4 separately assesses diachronic stability and the reliability of the system’s behavioral profile; they do not constitute additional hypotheses. K1 refers to the self-correction of norms and auditable correction of decisions. K2 denotes responsiveness to reasons that influence decisions. K3 concerns predictive reflexivity and the connection between deliberation and action. K4 denotes diachronic stability, that is, the durability of the decision profile and the possibility of reconstructing the history of changes. The chapter refers to James Moor’s typology of ethical agents and to Luciano Floridi and Jeff Sanders’s concept of Levels of Abstraction. The explicit-agent level provides a descriptive framework for the minimal threshold, but that typological classification does not replace demonstrating the capabilities required by H1 or satisfying the institutional conditions specified by WI. This approach eliminates both the excessive anthropomorphization of artificial intelligence and the reduction of responsibility to technical effectiveness.

Chapter 4 presents a pluralistic normative framework. The framework does not resolve the dispute between deontology, virtue ethics, and consequentialism, but orders their functions in the analysis of the actions of artificial intelligence systems. Deontology sets the limits of permissibility and protects the status of the addressees of norms. Virtue ethics, interpreted non-anthropomorphically, describes the stability and reliability of institutional practices in which artificial intelligence systems are designed, deployed, and supervised. Consequentialism resolves conflicts between reasons within the range of permissible variants of action. The dissertation proposes LexOrder: priority is given to the conditions of permissible action and normative addressability, followed by the stability and reliability of practices; only after these conditions are met are permissible options compared in terms of consequences and risk. FairAim complements this original proposal with a requirement for the impartial selection of principles and priorities.

PH1–PH3 provide the ethical and normative justification for the H1–H3 model but are not direct counterparts of the hypotheses. PH1 sets the limits of permissible action, whereas H1 specifies the capabilities required of a system; a permissible outcome does not, by itself, establish that H1 is satisfied. PH2 justifies the normative importance of stable and reliable practices; it is not the mechanism of functional affectivity specified by H2. PH3 orders conflicts between reasons within the range of permissible options but does not constitute the metacontrol required by H3.

Chapter 5 analyzes artificial subjecthood in classificatory and institutional terms. The distinction between an “agent” and a “subject” is crucial. An agent is a descriptive category referring to a system that exhibits interactivity, autonomy, and adaptability at a given Level of Abstraction, without any guarantee of normative status. A subject is a normative-institutional category; claims can be directed toward it, justifications can be expected from it, and responsibility can be enforced against it. This distinction shows that not every agent is a subject capable of bearing responsibility within institutional practices. The chapter presents a classificatory framework that links the ontological and functional dimensions of artificial subjecthood with conditions of intervention and interpretability. Requirements concerning meaningful human control and minimal interpretability are especially important. The analysis distinguishes the basic Delphi model, which predicts human moral judgments, from DelphiHYBRID, which combines predictions with symbolic rules. The explicitness of rules does not, by itself, establish that H1, F3, or H3 is satisfied. The conclusion of the chapter is cautious. An artificial intelligence system may function as a procedural quasi-subject only in a limited sense, after H1 and F3 have been demonstrated separately and appropriate R and I parameters have been ensured within WI. This does not mean granting full moral or legal personhood. On the contrary, it is meant to prevent responsibility from being removed from humans and institutions and sanctions from being shifted onto artifacts. Section 5.4 applies this framework to IDx-DR/LumineticsCore, classified as type I: A2/A4 × F2, a normatively saturated instrument within an accountable sociotechnical arrangement. The case distinguishes system auditability and institutional accountability from agential procedural addressability: satisfaction of H1 and H2 has not been established, and H3 remains unsatisfied considering the system’s documented functionality.

Chapter 6 examines the emulation, evolutionary, and revolutionary scenarios of artificial intelligence development. These scenarios are not treated as empirical forecasts, but as analytical

models that reveal different types of problems related to responsibility. The brain emulation scenario sharpens questions about continuity of identity, the multiplication of subjects, and the division of responsibility between emulated minds and the institutions that control them. The evolutionary scenario shows the gradual growth of complexity and the diffusion of responsibility in sociotechnical systems. The revolutionary scenario, associated with the idea of the singularity or intelligence explosion, functions as a warning boundary. It shows configurations in which the pace of system activity may exceed the capacities of human supervision, reflection, and correction. The chapter also develops an analysis of the responsibility gap. This gap may include the absence of a responsible subject but is not exhausted by that form: it is understood as a multidimensional violation of the conditions for attributing responsibility. The analysis uses qualitative metrics M1–M5, covering the identifiability of responsible actors, the auditability of reasons, reason-responsive control, the transparency and continuity of responsibility-attribution practices, and the relation between the pace of action or change and the time and knowledge available for reflection, control, and response. It distinguishes the capabilities of the technical system from the institutional conditions for applying the model; M4 concerns responsibility practices, whereas K4 assesses the stability of the system’s behavioral profile. The answer to the responsibility gap is not to grant legal personhood to artificial intelligence, but to design models of co-responsibility in which the roles of humans, institutions, and technical systems are explicit, subject to audit, and open to correction.

Chapter 7 synthesizes the results and conducts a philosophical verification of H1–H3 and WI. H1 is confirmed as a necessary condition for minimal procedural addressability. Without the ability to recognize reasons, disclose them, and subject decisions to correction, there is no basis for directing claims for justification toward a system. H1, however, is not sufficient for strong accountability. H2 is justified as a possible mechanism for stabilizing choices consistent with reasons but is not a necessary condition of the minimal threshold. Diachronic stability and the reliability of the system’s behavioral profile are assessed separately through K4; the result of that assessment does not, by itself, establish that H2 is satisfied. H3 is confirmed as the threshold of strong accountability, requiring metacognitive control, explicit resolution of conflicts between reasons, and the ability to present justification from a second-order perspective. WI completes the model by showing that even the fulfillment of H1–H3 is not sufficient without publicly

recognizable procedures of audit, supervision, correction, and attribution of responsibility to humans and institutions. The illustrative reference to IDx-DR/LumineticsCore shows that type I: A2/A4 $\times$ F2 may exhibit auditability and extensive institutional embedding without demonstrated satisfaction of the agential thresholds. This result illustrates the model's distinctions; it does not empirically validate the model.

The result of the dissertation is a hierarchy of the moral statuses of artificial intelligence, formulated as a two-track model. The axis of agency orders procedural roles from a morally significant artifact, through a normatively saturated instrument, to a procedural quasi-subject. The axis of vulnerability to harm orders the possible scope of protection from no special protection, through relational protection, to a precautionary approach under uncertainty about harm. This hierarchy does not lead to the automatic granting of rights or personhood to artificial intelligence. Rather, it allows the duties of humans and institutions toward systems, users, and people affected by algorithmic decisions to be defined more precisely. The author's contribution consists in proposing a threshold model of responsibility for artificial intelligence systems. This model distinguishes minimal procedural addressability from strong accountability, separates agential conditions from conditions of institutional application, and fills a gap in the understanding of responsibility as a gradual and multidimensional phenomenon. It preserves ontological minimalism without lowering normative requirements. Responsibility remains a practice of public justification, correction, and accountability, whose ultimate burden rests on humans and the institutions that organize the operation of artificial intelligence systems.

Keywords: artificial intelligence, responsibility, procedural addressability, accountability, moral subjecthood, responsibility gap, AI ethics, meaningful human control, interpretability.