AI & SOCIETY (2024) 39:1871–1882
https://doi.org/10.1007/s00146-023-01635-y

OPEN FORUM

Accountability in artificial intelligence: what it is and how it works

Claudio Novelli[1] · **Mariarosaria Taddeo**[2,3] · **Luciano Floridi**[1,2]

Received: 4 August 2022 / Accepted: 17 January 2023 / Published online: 7 February 2023

Abstract
Accountability is a cornerstone of the governance of artificial intelligence (AI). However, it is often defined too imprecisely because its multifaceted nature and the sociotechnical structure of AI systems imply a variety of values, practices, and measures to which accountability in AI can refer. We address this lack of clarity by defining accountability in terms of answerability, identifying three conditions of possibility (authority recognition, interrogation, and limitation of power), and an architecture of seven features (context, range, agent, forum, standards, process, and implications). We analyze this architecture through four accountability goals (compliance, report, oversight, and enforcement). We argue that these goals are often complementary and that policy-makers emphasize or prioritize some over others depending on the proactive or reactive use of accountability and the missions of AI governance.

Keywords Artificial intelligence · Accountability · AI Act · Governance · Policy

1 Introduction

Accountability is one of the cornerstones of the governance of artificial intelligence (AI). This is, among other reasons, because of the delegation of tasks (e.g., prediction or decision-making) to AI systems (henceforth AIs). Current AI policies, especially in the European context, acknowledge this aspect:

> "If we are increasingly going to use the assistance of or delegate decisions to AIs, we need to make sure these systems are fair in their impact on people's lives, that they are in line with values that should not be compromised and able to act accordingly, and that suitable accountability processes can ensure this".[1]

Despite its importance, accountability in AI is often defined too imprecisely, with undifferentiated references to the values, practices, and measures it encompasses. This is due to the multifaceted nature of accountability—which is a context-dependent relation (Sinclair 1995), the inherent ambiguity of political processes (Olsen 2017), and to the sociotechnical structure of AIs (Theodorou and Dignum 2020). In sociotechnical systems, rules and customs of different contexts are intertwined, and, as we shall see, this has major implications for accountability.

Unfortunately, an imprecise definition of accountability is problematic, not least because it risks undermining the public debate and policy-making. This does not happen often, especially where laws exist to define and ascribe accountability. But when it does, typically where regulations are less developed, an imprecise definition of accountability hides the implicit trade-offs among different political choices over which accountability regime should be enforced. This is problematic especially when political and legislative agreement have not yet been formed, including the accountability of many AI services. In this article, we address this lack of clarity by analyzing the concept of accountability in AI and defining its features and goals.

The article is structured as follows. Section 2 defines accountability as a specific relation of *answerability*, and identifies necessary conditions of possibility: *authority recognition*, *interrogation*, and *limitation of power*. Section 3

✉ Claudio Novelli
claudio.novelli@unibo.it

[1] Department of Legal Studies, University of Bologna, Via Zamboni, 27, 40126 Bologna, IT, Italy

[2] Oxford Internet Institute, University of Oxford, 1 St Giles', Oxford OX1 3JS, UK

[3] Alan Turing Institute, British Library, 96 Euston Rd, London NW1 2DB, UK

[1] AI HLEG, European Commission (2019) A definition of AI: main capabilities and disciplines. https://ec.europa.eu/digital-inglemarket/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines.

analyzes the "architecture" of accountability in terms of features: *context, range, agent, forum, standards, process,* and *implications*. The identified features include several values, practices and measures and Sect. 4 discusses them in terms of accountability goals: *compliance*, *report*, *oversight*, and *enforcement*. We argue that, although these goals are often complementary, policy-makers tend to emphasize or prioritize some of them over others. This general analysis of accountability is then applied to AI. Section 5 clarifies that providing an "architecture" of accountability in AI also requires taking a sociotechnical approach. Section 6 then shows how the seven features identified in Sect. 3 characterize accountability in AI. Section 7 further analyzes and restricts the content of the accountability relation in AI through the accountability goals described in Sect. 4. Section 8 shows two factors that lead, or should lead, AI policy-makers to emphasize some goals over others, namely the use of accountability proactively or reactively and the governance missions in AI. Section 9 concludes the article.

1.1 The state of the art: accountability in European regulations

Accountability is often broadly defined. This is clear in some of the major European documents on AI. Let us consider the High-Level Expert Group (HLEG) reports, the GDPR and the Artificial Intelligence Act (AIA).

In the HLEG reports, accountability is defined both as a principle that ensures compliance with the key requirements for a trustworthy AI—in this sense, it works as a *meta-principle* (Durante and Floridi 2022)—and as a set of practices and measures, e.g., audit, risk management, and redress for adverse impact. The polysemic nature of accountability is confirmed in the Assessment List for Trustworthy Artificial Intelligence (ALTAI) by the same expert group:

> "This term refers to the idea that one is responsible for their action—and as a corollary their consequences—and must be able to explain their aims, motivations, and reasons. Accountability has several dimensions [and] might also express an ethical standard, and fall short of legal consequences […]"[2]

Similar considerations apply to the concept of accountability advocated in the GDPR (e.g., Articles 5 and 24). Here, accountability works as a meta-principle directed at data controllers so that they demonstrate, by virtue of their information background, compliance with GDRP requirements in the processing of personal data and as a remedy mechanism for failure to comply with them:

> "The controller shall be responsible for, and be able to demonstrate compliance with [fairness, transparency, purpose limitation, data minimisation, storage limitation, accuracy, confidentiality, etc.]"[3]

Finally, the AIA also contains an undefined concept of accountability, aligned with the risk-based regulatory approach: providers and implementers of AIs are accountable for different reasons and in different ways depending on the risk level of the respective AIs.

2 What is accountability and what does it require? Accountability as an answerability relation

Accountability has many definitions but, at its core, is an obligation to inform about, and justify one's conduct to an authority (Bovens 2007; Lindberg 2013; Mulgan 2000; Thynne and Goldring 1987). More formally, accountability denotes a relation between an *agent A* and (what is usually called) a *forum F,* such that *A* must justify *A*'s conduct to *F*, and *F* supervises, asks questions to, and passes judgment on *A* on the basis of such justification (Bovens 2007, 450). Both *A* and *F* need not be natural, individual persons, and may be groups or legal persons.

Understood as a relation, accountability often counterbalances another relation that logically precedes it: the (potentially implicit) *delegation* to *A* of some tasks *T* (actions, services, or powers to act etc.) by a source *P*, (usually called) *the principal*, on behalf of which *A* acts (Lindberg 2013; Mulgan 2003). Note that *P* and *F* may differ: *A* may be delegated *T* by *P* but be accountable for *T* to F. Thus, accountability can be a binary or ternary relation. Since our interest in this article is to understand accountability itself, in what follows we shall assume that $F = P$, and hence use *forum* and *principal* interchangeably, depending on the relevant context. Nothing in this article depends for its cogency on this assumption.[4]

Once it is defined in the previous terms, accountability relation can be modeled in terms of content, as *answerability* (Akpanuko and Asogwa 2013; Olsen 2017). As such, it has the following necessary conditions of possibility: *authority recognition*, *interrogation*, and *limitation of power*. They are intertwined but lets us discuss them separately.

Authority recognition results from the delegation of tasks and is mutual: *P* (e.g., citizen) grants *A* (e.g., public servant) authority to serve its interests, while *A* recognizes that *F*

[2] Accountability in the Glossary of Assessment List for Trustworthy Artificial Intelligence (ALTAI).

[3] Art. 5(1) and (2) of the GDPR.

[4] However, there are differences between the principal and the forum, e.g., the principal typically sets the conditions and rules of the accountability mechanism, while the forum merely implements them.

Table 1 Accountability features

Features	Explanations
1. Context (what for?)	Fields in which an accountability relation is established
2. Range (about what?)	Tasks, like actions, services, decisions, and assessments taken by the accountable agent
3. Agent (who?)	The entity who exercises the delegated powers, accepting to be blamed or praised
4. Forum (to whom?)	The entity engaged in actual interrogation and supervision and/or the bearer of the interests served through delegation of tasks (*principal*)
5. Standard (according to what?)	Principles, rules, and benchmarks against which the conduct of the accountable agent is assessed
6. Process (how?)	Procedures through which the agent is called to account
7. Implications (what follows?)	Consequences, formal or informal, triggered by the accountability assessment

(which could be identical to *P*) will hold it accountable for the way such interests have been served. Without authority recognition, accountability would become mere "grace-and-favor reporting or informing" (Mulgan 2003, 11). However, recognition of the authority alone is not sufficient, because the legitimacy of standards and procedures of the accountability relation also needs to be recognized (Grant and Keohane 2005).

Interrogation refers to the fact that *A* is exposed to *F*'s scrutiny. If this were not the case, the traditional distinction between 'accountability' and 'moral responsibility' would collapse. Indeed, while accountability presupposes scrutiny from an external point of view, moral responsibility may refer only to an internal point of view: a personal judgment of one's agency, presupposing "the capacity to act from free choice and with due concern for one's duties and obligations" (Mulgan 2003, 15).[5] To be accountable only to oneself means not being accountable at all.

Limitation of power constrains the arbitrary exercise of delegated *T* by allowing *F* to monitor the *A*'s performance and evaluate its results (Lindberg 2013; Mulgan 2003). In democratic societies, accountability, as an effective constraint on delegated powers, is a prerequisite for the legitimacy of these powers themselves. Indeed, the very definition of democracy may be grounded on accountability: "Modern political democracy is a system of governance in which rulers are held accountable for their actions in the public realm by citizens […]" (Schmitter and Karl 1991, 76).

Understanding accountability in terms of answerability and its three necessary conditions of possibility is a step forward but more analysis is needed to grasp how answerability is structured and applied. This is the task of the next section.

3 What does accountability include? The features of accountability relation

To understand what the relation of answerability includes, one needs to consider its (1) context (what for?); (2) range (about what?); (3) agent (who is accountable?); (4) forum (to whom an account is due?); (5) standards (according to what?); (6) process (how?); and (7) implications (what follows?).[6]

Table 1 Provides a synthetic overview.

Logical priority is given to the context, because without it one could not identify the agent, the forum, and the content of the accountability relation.

Table 2 illustrates some typical cases of accountability: electoral, juridical, and administrative, based on (Bovens 2007; Mashaw 2006).[7]

To exemplify: electoral accountability (what for?) for choices on the recruitment of representatives (about what?)

[5] In many cases accountability presupposes moral responsibility: the subject called upon to answer under criminal law is also the one who can be internally responsible. Yet, one can be accountable without being morally responsible, as in the case of collective actions, disclosed agency, or strict liability. In the opposite case, one can be internally responsible without being accountable to anyone specifically (Schedler 1999, 19). Thus, the "division of linguistic labor" (Mulgan 2003, 17) between the two concepts, and then the assimilation of accountability with answerability, has a positive explanatory function.

[6] A similar approach is promoted by Mulgan, who however circumscribes these aspects to four dimensions (who? to whom? for what? how?) (Mulgan 2003).

[7] Other taxonomies consider different aspects, e.g., source and degree of supervision, or spatial direction, contrasting upward and downward accountability (Lindberg 2013; Romzek and Dubnick 1987; Schedler 1999).

Table 2 Examples of accountability types

Features	Examples		
Context (what for?)	Electoral	Juridical	Administrative
Range (about what?)	Choices of political direction, laws, and recruitment	Conducts, omissions, and decisions	Policy implementation
Agent (who?)	Representatives, leaders, parties, governments, and institutional bodies	Natural persons, legal persons, states, and assets	Public officials and institutions
Forum (to whom?)	Citizens, voters, taxpayers, political parties, and institutions	Individual and collective entities (including states and courts)	Citizens, auditors, inspectors, and ombudsman
Standard (according to what?)	Reliability, coherence, and ideology	Legal rules, principles, and precedents	Efficiency, effectiveness, and legal norms
Process (how?)	Public debate (media), internal or external vigilance (e.g., judicial review), and elections	Judicial and extra-judicial review	Auditing, internal supervision, and judicial review
Implications (what follows?)	Electoral outcomes, political reputation, careers, and funding	Reparations, remands, detentions, fines, and prohibitions	Certifications, validations, revocations, penalties, suspensions, and seizures

by political leaders (who?) to the electorate (to whom?) is assessed against the standard of reliability (according to what?) through elections (how?) and can imply electoral failure (what follows?).

Table 2 includes several cases, it enables the reader to consider alternative and intersecting taxonomies. For example, depending on the nature of the agent (who?), one may distinguish between individual, corporate, collective or hierarchical types of accountability (Bovens 2007, 461). Also, depending on the nature of the range (about what?), one may distinguish procedural and outcome accountabilities. To structure the accountability relation in AI more comprehensively, we shall integrate these additional classification criteria in Table 3 (see Sect. 6).

4 What goals does accountability serve? Compliance, report, oversight, and enforcement

The features listed in Table 2 include several values, practices and measures to which accountability can refer. This wide content is often[8] filtered and restricted considering the *goals* that accountability is supposed to serve. In this section, we identify four goals which are widely acknowledged (see references below) to shape the way policy-makers envisage accountability regimes in governance frameworks: *compliance*, *report*, *oversight* and *enforcement*. They are introduced in a loosely logical order.

(1) *Compliance*. The goal is to bind the agent to align with ethical and legal standards. From this stance, as Bovens points out, accountability: "is used as a normative concept, as a set of standards for the behavior of actors, or as a desirable state of affairs […] Accountability in this very broad sense […] comes close to 'responsiveness' and 'a sense of responsibility', a willingness to act in a transparent, fair, and equitable way" (Bovens 2010, 949).

(2) *Report*. The goal is to ensure that the agent's conduct is properly recorded to explain and justify it to the forum (or the principal). The reporting of relevant information enables the forum (or the principal) to challenge and disapprove the agent's conduct. Determining which information is relevant is not always easy, but can be based on the requirements of the associated oversight: "In many instances [report] is a mirror of (and surrogate for) the act of direct monitoring by a principal of the behavior and act" (Dubnick 2005, 383).

(3) *Oversight*. The goal is to examine information, obtain evidence, and evaluate the agent's conduct. Oversight must allow for solid scrutiny, also in the form of ex-ante control of decision-making processes by the forum. Ex-post oversight assesses whether explanations and justifications are acceptable to the rules of the deployment context, e.g., judicial review.

(4) *Enforcement*. The goal is to determine what consequences the agent must bear—e.g., sanctions, authorisations or prohibitions—according to the evidence gathered during the report and oversight.

[8] These goals, though differently named, emerge in the literature. In political accountability, Schedler recognizes three "dimensions": information, justification, and punishment (Schedler 1999). Mulgan talks about three "stages": information, discussion, and rectification (Mulgan 2003). Rubenstein considers three "parts" of the accountability process: standard-setting, information gathering and the imposition of sanctions (Rubenstein 2007). Bovens, finally, considers: information, debate (or interrogation), and judgment (Bovens et al., 2008).

Policy-makers can pursue the previous accountability goals disjointedly or simultaneously (Schedler 1999, 17).[9] Yet, even in the latter case, there is a descriptive and normative reason for keeping goals separate. In areas where accountability is more legislated, policy-makers usually emphasize or prioritize some of these goals over others (Srinivasan and San Miguel González 2022). And where accountability legislation is less developed, we suggest that policy-makers *should* emphasize or prioritize some goals over others depending on specific governance factors. We shall return to this point in Sect. 8 in relation to the governance of AI. Here, let us close with a brief discussion of the two rationales.

Regarding the descriptive rationale, although they may be complementary, policy-makers often highlight only some of the accountability goals, presupposing others or delegating their elaboration and pursue to other subjects and places. The absence of one of these four goals does not invalidate accountability regimes but generates surrogates (Rubenstein 2007). For example, when accountability is primarily an enforcement tool:

> "We talk of people being 'accountable' or 'answerable' to other people and mean nothing more by it than that the people to whom there is accountability are in a position to inflict punishment on those being held accountable should they deem them guilty of misconduct […] what might be called a 'coercive' or rather 'purely coercive' variety that can operate quite independently of the informative and has, if anything, a better claim to the title of 'accountability'" (Kaler 2002, 329).

In other cases, the report is disregarded due to scarce and unreliable information, e.g., in electoral accountability. Also, when it is not possible to impose sanctions without a centralized government, accountability is decoupled from enforcement, e.g., in global accountability (Grant and Keohane 2005). At the same time, it is possible to overlook enforcement as analogous remedies can already be taken during oversight, e.g., when ombudsmen monitor public officials' performance and provide recommendations for improvement (Mulgan 2003).

Regarding the normative rationale, where regulation and debate on accountability is less developed, undifferentiated use of accountability may hinder political coordination and cause regulative noise, whereas a goal-based analysis enables policy-makers to adopt the most suited accountability regime as opposed to a governance framework.

So far, we have provided a general analysis of accountability: conditions of possibility (Sect. 2), architecture (Sect. 3) and goals (Sect. 4). This analysis yields a picture of accountability sufficiently specific to apply it to AI. We shall do so in the next section.

5 Accountability in AI: a sociotechnical approach

Generally speaking, accountability in AI relates to the expectation that designers, developers, and deployers will comply with standards and legislation to ensure the proper functioning of AIs during their lifecycle (Fjeld et al. 2020).[10] This is the context (see Table 2).

However, AIs are neither mere artifacts nor traditional social systems: technological properties often make the outcome of AIs opaque and unpredictable, hindering the detection of causes and reasons for unintended outcomes (Tsamados et al. 2022). Various factors lead to (uses of) AIs perpetrating wrongdoings, consider for example the case of AI perpetrating undue discrimination, this can result from biased training data, system bugs, programmer errors, misuses, or the replication of social discrimination; and sometimes a combination of these factors. The nature of AIs makes it problematic to assess accountability for such outcomes. This is because opaque and unpredictable outcomes of AIs have similar consequences to the 'many hands' problem (Cooper et al. 2022; Thompson 1980): the impossibility of pinpointing individual responsibilities in systems that involve multiple actors and resources. This creates suboptimal equilibria in which distributed responsibly (Floridi 2013, 2016) means that nobody may feel oblige to prevent negative consequences (Hardin 1968). Moreover, the symmetrical problem of 'many eyes' can arise, that is, a multiplication of fora, each with different expectations and judgment criteria (Bovens 2007). Poor administration of these problems causes two opposite effects: *accountability gaps*, where no one is held accountable, or *accountability surpluses*, where procedures are inefficiently accumulated (Bovens 2007; Busuioc 2021). In both cases, this results in the lack of virtuous practices to mitigate the risks of undesirable outcomes and of effective redress for victims.

Technological and organizational peculiarities of AIs call for a sociotechnical approach to accountability, a point

[9] The content of accountability relation depends on how these goals are pursued, generating different tasks and obligations for the forum (or the principal) and the agent. However, even though policy-makers may pursue the goals disjointedly, from the viewpoint of the forum (or the principal) and the agent these goals tend to be unitary.

[10] For a comprehensive definition of AIs, see the one provided by the expert group appointed by the European Commission in 2019 (AI HLEG) in the 'Ethics guidelines for trustworthy AI'.

insufficiently addressed in the relevant debate. As Theodorou and Dignum point out:

> "[…] technology, or the artefact that embeds that technology, cannot be separated from the socio-technical system of which it is a component. This system includes people and organizations in many different roles (for example, developer, manufacturer, user, bystander or policymaker), their interactions and the processes that organize these interactions" (Theodorou and Dignum 2020, 10).

The notion of sociotechnical systems was developed during the 1960s by the Tavistock Institute for work organization studies (e.g., factory work). It refers to complex hybrid systems in which human and technical resources are joined in goal-directed behavior (Baxter and Sommerville 2011; Long 2013).[11] The performance of a sociotechnical system relies on the joint optimization of tools, machinery, infrastructure and technology (e.g., software), on the technical side, and of rules, procedures, metrics, roles, expectations, cultural background, and coordination mechanisms, on the social side. The interplay between these components prevents their disentanglement as single observable parts for specific outcomes, just as it prevents the detection of a general function, as such hybrid systems are embedded in a network of individual actions.

Against this background, accountability for AIs' behavior should be distributed according to the network of technological and social inputs (Kaminski 2020). From a theoretical perspective, a sociotechnical picture of accountability in AI helps to unpack three features of its architecture (see Table 3, next section):

- the *range* (about what?): this is important because pre-existing values, rules, habits and incentives of a specific social environment affect task execution by AIs, and vice versa (Selbst et al. 2019), possibly leading to unintended events;
- the *agents* (who?): this requires figuring out the interaction among technology, humans, and the environment, e.g., the division of cognitive labor between AIs and human agents, the level of autonomy of the AIs, legal constraints, task distribution among workers, and procedures (Zweig and Raudonat 2022);
- the *standards* (according to what?): because the agent may have to provide pieces of evidence, explanations and justifications to objections on both technical and social sources of conduct (Binns 2018).

[11] One way to see the difference between these components is that the technical components are governed by natural laws but are insufficient to explain the social components (Vermaas et al. 2011).

By extension, it will be easier to identify the remaining three features of the accountability architecture in AI: a socio-technical picture of range and agents facilitate the identification of the *forum* (to whom?), while standards facilitate the identification of *processes* (how?) and *implications* (what follows?).

6 The features of accountability in AI

In this section, we apply the seven features mentioned in Sect. 3 to accountability in AI. Each feature has sub-features that illustrate the variety of values, practices, and measures to which accountability in AI can refer.

The *context* (what for?) of accountability of AIs can be identified by the field of use, by the function, or by the level of autonomy of the AIs in question.[12] An intersection of these criteria specifies the context, e.g., machine vision applied to medical diagnosis for decision support.

The *range* (about what?) can be defined around the three sets of tasks of an AIs' lifecycle: design, development or deployment. Design tasks mainly concern planning, e.g., the choice of technology and infrastructure, interface design, data and development strategy. Development tasks involve programming, training, engineering, and testing AIs. Finally, deployment tasks concern using, monitoring and maintaining AIs, according to the rules of a specific context (Desouza et al. 2019).

Design, development, and deployment are performed by different *agents* (who?). They can be identified individually, corporately (e.g., as detached legal entities), collectively (all accounting in the same way), or hierarchically (according to their roles and functions).[13] The same applies to the *forum* (to whom?). For example, it may be defined collectively, as subjects bearing with the consequences of the actions performed or mediated by AIs (i.e., decision-subjects), or subjects whose data are used to train the AIs (i.e., data-subjects), or shareholders, or a domain practitioners (e.g., clinicians).[14]

[12] Classifications 'by function' and 'level of autonomy' can be seen as subsets of that by context, but it is often the case that accountability issues in AI are regulated by direct reference to functions and autonomy.

[13] Agents might also be identified by the type of unintended events: i.e., mistakes, misuses or accidents. The most problematic ones are 'Accidents', caused by the autonomous agency of AIs and not by coders' inputs as responses to particular situations. In such cases, liability might be shared between developers/producers of AIs (Martin 2019) or managed through insurance mechanisms (Zech 2021).

[14] Domain practitioners are all those parties who handle and take responsibility for the selection, procurement or application of AIs in their specific domain (Barclay & Abramson 2021).

Tasks included in the accountability *range* must be assessed under different *standards* (according to what?). Indeed, contention can take place on both normative and epistemic grounds (Binns 2018), concerning legal rules, ethical principles, or technological requirements of AIs.[15] Note that the boundaries between these standards are often blurred: e.g., privacy is a legal notion, but it can be seen as an ethical and also a technological standard.

The adherence to these standards is assessed through rules, metrics and procedures. This is the accountability *process* (how?). This process can be driven by creators of AIs (e.g., internal supervision), by third parties (e.g., external audit) or, at least partially, by human–machine interaction (HMI). The AI Act, for instance, requires internal or external conformity assessment procedures depending on the type of high-risk AIs. AIs used in the administration of justice and democratic processes requires only an internal conformity assessment (Art.19), but for AIs used for biometric identification, and without harmonized standards, the conformity assessment requires the involvement of an external notified body (Art. 30 ff.). The AI Act also assigns a key role to human–machine interface tools, through which natural persons can oversee high-risk systems during their use (Art. 14).

The last feature is *implications* (what follows?). The debate on which consequences, formal or informal, should follow the accountability assessment in AI is ongoing. Different implications follow unlawful facts, e.g., harmful or illicit conduct (whether intended or unintended), lawful facts, e.g., the proper functioning of the system, or mere decisions that do not yet produce effects, e.g., data strategy. Using an example to illustrate the previous analysis, in granting a loan (what for?) a bank using an AI (who?) can be made accountable to a customer (to whom?) for the creditworthiness evaluation algorithm (about what?) due to discriminatory data (according to what?), by requesting explanation (how?) and possibly getting a revision (what follows?). Table 3 illustrates the content of these features (and sub-features).[16]

7 The goals of accountability in AI

Applying to AIs the analysis developed in Sect. 4, we now obtain the following clarifications:

(1) *Compliance* is about binding AIs to align with ethical, legal, or technical norms. This goal defines the design, development, and deployment standards to be met throughout the entire lifecycle of an AIs, but it is rather generic if it is not implemented by good practices. Compliance is often translated into preliminary checks by AIs providers, as is the case in the AI Act where ex-ante compliance is crucial to bring high-risk AIs to market.

(2) *Report* represents the dialogical dimension of accountability: practices ensuring explanation and justification of AIs' behaviors. For example, it protects the right to object to automated decisions, as also provided for in Articles 21 and 22 of the GDPR. As AIs are frequently opaque (black-boxed) report as mere transparency or complete explanation have been replaced by more functional approaches, like explainable and interpretable AI.[17] An explainable AIs does not report all available information, but only those conductive to contextual explanations, i.e., compared to counterfactual cases and relevant to the interaction between the agent and the forum (Miller 2019). An interpretable AIs describes its internals in a user-understandable way, thus sacrificing completeness to meet the user's cognition, knowledge, and biases (Gilpin et al. 2018).[18]

(3) *Oversight* seeks to find relevant facts or information, and create evidence, to evaluate the life-cycle performance of AIs. Oversight has become essential for AI governance, as it is also stressed by the AI Act (e.g., article 14). It can be carried out by different bodies, internal or external to the organization, or through human–machine interfaces. In the latter case, oversight is enabled by the design of the system itself, before it is put on the market and operated (Kroll 2020). Overseers may act at different levels that may overlap, e.g., an internal audit is compatible with judicial review.

(4) *Enforcement* ties the monitoring and evaluation of the performance of AIs to formal or informal consequences. This is also with an aim of deterring unwanted behaviors. In the case of AIs, enforcement can consist of either the outcome of conformity assessment (art. 43 AI Act), i.e., fees, or the outcome of a proper judicial review, e.g., compensation for damages.

As mentioned in Sect. 4, these goals can be pursued disjointedly or with a different emphasis when policy-makers

[15] These criteria often intersect, so that elements of technological robustness become ethical principles of AI (in each case they are strongly functional to the fulfillment of ethical principles).

[16] The table does not cover all the criteria for ranking or ordering values, practices and measures included in the accountability relation.

[17] Literature on explainable AI can be influenced by sociotechnical approaches (Ehsan et al. 2021).

[18] However, AIs should not only persuade the user and the trade-off between interpretability and completeness needs to be carefully evaluated.

Table 3 Accountability features in AI

Features	Sub-features		
Context (what for?)	[AIs by field] finance, healthcare, justice, military, commerce, engineering, automotive, and public administration	[AIs by function] natural language processing, machine vision, information retrieval, filtering, classification, and robot control	[AIs by level of autonomy] manual control, action support, shared control, decision support, blended decision-making, automated decision-making, and full automation[b]
Range (about what?)	[Design] planning, audience focus, architecture design, data strategy, development strategy, and interfaces design	[Development] coding, implementation, model training (e.g., data processing), security mechanisms (IP protection), testing, and integration	[Deployment] monitoring, maintenance and use
Agent (who?)	[Individuals] AI designer, data experts, AI developers, manufacturers, and domain practitioners	[Hierarchies] patent-holders, managers, superiors, supervisors, and testers	[Corporates or Collectives] policy-makers, development firms, data controllers, and shareholders
Forum (to whom?)	[Individuals] decision-subjects, data-subjects, and domain practitioners (e.g., customers and users)	[Hierarchies] managers, superiors, and supervisors	[Corporates or Collectives] citizens, shareholders, external bodies, authorities, and institutions
Standard (according to what?)	[Legal] torts, crimes, unfair commercial practices, privacy, and risk tolerance (e.g., AI Act)	[Ethical] fairness, transparency, human autonomy, inclusion, vulnerability, trustworthiness, and sustainability	[Technical] robustness, adaptability, accuracy, efficiency, maintainability, (cyber)security, and self-healing
Process (how?)	[Internal] internal audits, simulations, self-assessments[a], and post-market monitoring	[HMI] feedback loops, supervisory controls, and interactive machine learning	[External] systems validations, external audits, external conformity and impact assessments[c], ombudsman, and judicial reviews
Implications (what follows?)	[Decisions] recommendations, approvals, refusals, and prohibitions	[Lawful facts] reputation, market profit or loss	[Unlawful facts] reparations, remands, fines, disciplinary measures, detentions, suspensions, revisions, and revocations

This framework is intended to be illustrative only, as alternative sub-features can be considered.

[a]Consider the European Assessment List for Trustworthy AI (ALTAI) for self-assessment (2020), a checklist for AI development and deployment. Likewise, the Algorithmic Accountability Act (AAA) of 2022 presented in the US Congress requires companies to assess the social impact of their automated decision systems

[b]These autonomy levels refer to the taxonomy developed by Endsley and Kaber (Endsley and Kaber 1999)

[c]Art 31 ff. AI Act. However, for certain types of high-risk AIs, these assessments can also be internally-driven

regulate accountability in a given area. The novelty of the AIs, at least from the regulatory point of view, often pushes prioritizing the definition of common standards (i.e., compliance goal). Indeed, in this context accountability is often referred to as a general principle or a quality that AIs should have, rather than a concrete legal mechanism.

A goal-based analysis of accountability is beneficial as each goal relates to different sociotechnical aspects of AIs, duties of the policy-makers, and trade-offs between values and interests of stakeholders. Three points are worth highlighting.

First, accountability goals have different regulative focuses within sociotechnical systems. For example, if accountability is primarily aimed at compliance, the regulative focus should be on the duties of designers and developers to build AIs that meet ethical, legal, and technical standards. Whereas, if report is prioritized, the focus is on information exchange or transparency requirements (e.g., about data sources, metrics, or procedures), whether the goal is oversight, competencies and powers of the overseer or the jurisdiction of the court should be of major concern. Finally, relevant aspects for the goal of enforcement are those related to the duration and gravity of the infringement (e.g., Art. 72 AI Act) or the distribution of liabilities among the stakeholders of the sociotechnical system.

Second, distinguishing accountability goals enables one to identify different degrees of commitment for policy-makers and those who enact and execute the accountability regimes, e.g., subsidiary lawmakers, national authorities, and judges. Compliance, being value-based, leaves greater discretion to subsidiary lawmakers to design accountability regimes according to their preferences and needs. This is not the case with regulatory "richer" goals, like oversight, which is procedure-based. This is crucial where different legal systems interact, not only internationally but also within the EU.

Third, the goal-based analysis highlights the spectrum of trade-offs between interests that inform political agreements on accountability regimes. The political process leading to an accountability regime in which, for instance, compliance prevails is different from one in which enforcement prevails. In the EU, while consensus on AI compliance is consolidating around general values—especially fairness, trustworthiness, and privacy—oversight or enforcement rules are still quite undefined and left to political and legal settlements of the EU Member States.

The last question to be addressed in this article is how policy-makers choose, or *should* choose, one of the four goals over the others. In the next section, we identify two factors.

8 What determines the choice of goals? The governance behind accountability in AI

Preferences for specific goals are determined by the *proactive* or *reactive* use of accountability and the AI governance *missions*.

Proactive accountability sets the agenda, specifying the requirements that make an AIs accountable before the occurrence of any event. In this sense, accountability answers the question: '*what should an accountable AIs look like?*'. Reactive accountability concerns implementation, that is, the sequence of measures triggered by the relevant event. The question it answers is: '*how should (the use of) an AIs be held accountable?*'. Both uses are defined upstream, but while proactive accountability is implemented ex-ante, reactive accountability is implemented ex-post.

As shown in Fig. 1, proactive accountability focuses on generating intended outcomes and preventing or reducing unintended ones. It aims to identify and correct organizational aspects that cause misconduct and mistakes before these occur. Proactive accountability requires building AIs with clear goals and accurate division of roles, responsibilities, and lines of command. From this perspective, accountability is a *virtue* that AIs must acquire, rather than just a *mechanism*.[19] To correct systemic distortions from a proactive stance, greater emphasis is given to standard-setting (compliance) and preliminary checks on the conformity and the impact of AIs (oversight). Less attention is paid to enforcement, which is often triggered by unwanted events showing the limits of proactive accountability. The report goal has a proactive function when used to empower consumers to make informed choices about AIs. An example of a "proactive report" is that promoted by the 2022 US Algorithmic Accountability Act (AAA), which requires the Federal Trade Commission to publish a repository on critical decisions made by automated decision-making systems.

Reactive accountability triggers rewards or sanctions once the event has occurred. When the event is undesirable, reactive accountability aims to redress the effects of failures. From this view, accountability is much more a mechanism than a virtue. Greater emphasis is given to report and enforcement: after the event, emphasis is given to the explanations and justifications to be provided by the agent to the forum (report) and the consequences of the accountability assessment (enforcement). Oversight has also an important role here, but as a retrospective review (e.g., judicial review) rather than as a preliminary check.

Proactive and reactive accountability are often combined, but generally one is more prominent. The AIA, for example,

[19] The distinction between accountability as a *virtue* and as a *mechanism* is by Bovens (Bovens 2010).

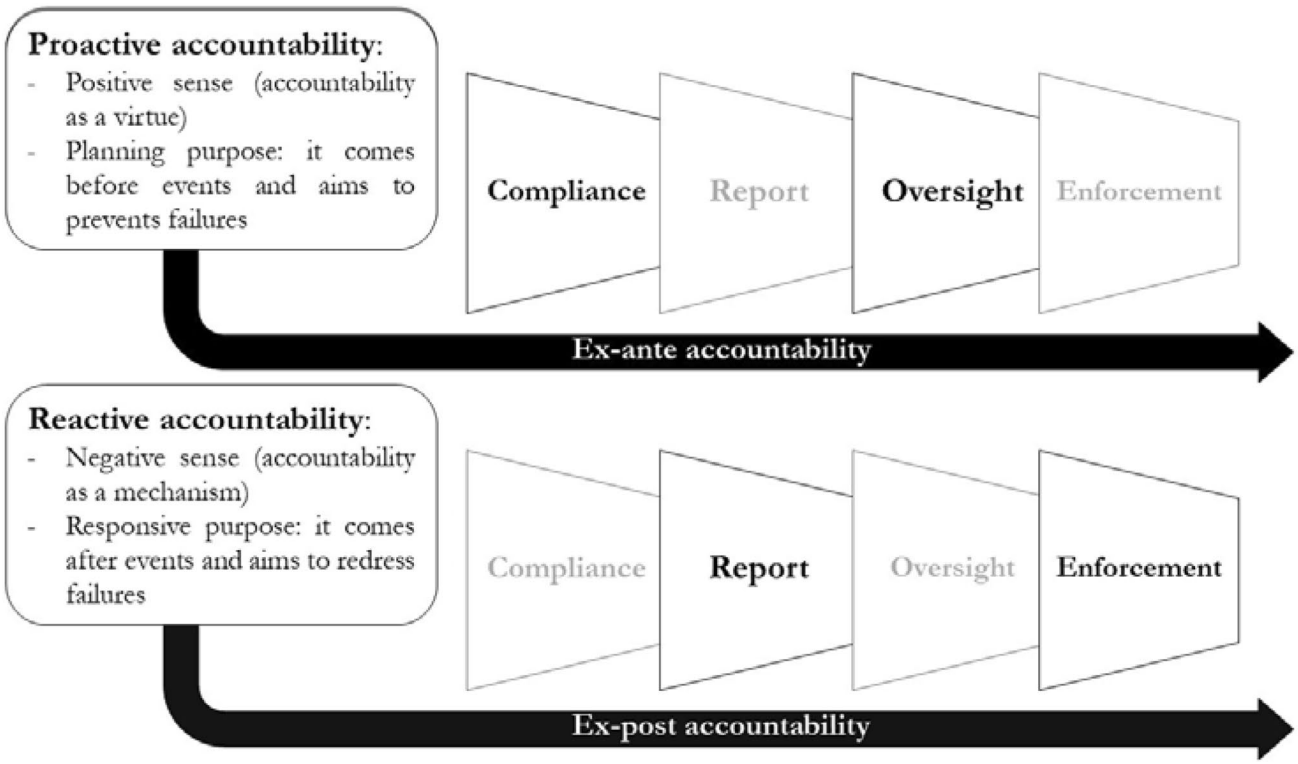

Fig. 1 Proactive and reactive accountability

combines both, but proactive accountability prevails as the regulatory burdens on AIs' providers regarding compliance and oversight are prioritized, especially in arts. 9 and 17. This is also the case in the HLEG report on Ethics guidelines for trustworthy AI (2019), which informed the AIA.[20] Overall, at least in the EU, reactive versions of accountability in AI are likely to prevail as general principles and guidelines are implemented into specific legal frameworks.

These accountability uses alone do not explain yet the social and political reasons that induce, or should induce, policy-makers to emphasize or prioritize some goals over others. This requires looking at the broader governance of AI, and its missions, behind accountability regimes. Traditionally, accountability policies serve different governance missions, all compatible and relevant to AIs.

One of the main missions of AI governance is *value harmonization* for design, development, and deployment of AI technologies. As accountability regimes govern moral expectations and claims in social interactions, they promote alignment on specific values and beliefs (Dubnick 2011). In AI, accountability fosters convergence toward values like fairness, inclusivity, and transparency. Under this governance mission, the goal of compliance should be emphasized. Value harmonization engenders trust (Taddeo 2017), which is another key mission of AI governance, one for which accountability plays a key role.

Good governance ensures that expectations about the benefits of AI technologies are met. This requires understanding technological potential and risks, building an appropriate infrastructure to integrate AIs into specific environments, and developing safe, robust, and efficient AIs. Accountability rules, both proactively and reactively, introduce oversight and transparency mechanisms throughout the lifecycle of AIs, and thus play a central role in *fostering (public) trust* in AI (Fjeld et al. 2020). The greater the legal certainty on accountability schemes, the greater the public and private trust in AIs. AI market growth comes through legal certainty and widespread trust. Under this governance mission, oversight and report should be emphasized.

To drive its impact on society, governance policies must distribute the costs and risks of AI. An efficient *allocation of social costs* should incentivize technological innovation, prevent or minimize damage, and ensure redress for victims. As accountability regimes assign duties and liabilities for social activities, they also distribute negative externalities. There will be different social effects depending on how the accountability burdens are distributed among stakeholders of AIs. The AI Act, for example, places most of the

[20] Generally speaking, in AI policy documents where accountability appears as an ethical principle, a proactive version is being used.

accountability burdens on developers and providers of high-risk AIs.[21] Under this governance mission, the focus should be on enforcement.

Other governance factors are relevant for framing accountability in AI, e.g., whether missions are long-term or near-term, whether policies are intended to be flexible or rigid, and the choice of the legal instrument to carry out them. The choice of policy-makers as to which values, practices, and measures should be emphasized in concrete cases will be more effective if the uses of accountability (whether proactive or reactive) and goals are consistent with the governance background. For overarching missions, like setting the political agenda, an accountability regime focused on oversight or enforcement may not promote adequate public debate or accurate regulation. Conversely, accountability focused on compliance or report may be insufficient to pursue narrower missions, like providing sector-specific practical guidelines on AIs. These governance missions frequently coexist, albeit with different weights in individual policies or regulations of AI.

9 Conclusions

Addressing accountability in AI requires tackling several difficulties, like the broad definition of accountability and the opacity of AIs. In this article, we defined accountability as a relation of answerability requiring authority recognition, interrogation and limitation of power. We then specified the content of the answerability relation through seven features: context, range, agent, forum, standard, process, and implications. We discussed the set of values, practices, and measures included in this "architecture" in terms of goals that accountability may serve in a governance framework, these are: compliance, report, oversight, and enforcement. In the second part of the article, we applied this analysis of the structure and content of the accountability relation to AI. We adopted a goal-based analysis as a useful guide to examine policy strategies for AI. We contextualized accountability within a broader institutional scenario, identifying two factors that do or should determine the choice of policy-makers when regulating AI: a proactive or reactive use of the accountability relation and the governance objectives. It seems obvious and yet inevitable to conclude that the balance that needs to be struck between different accountability policies, their specific formulation and implementations will remain a matter of ethical, legal, and political deliberation about preferred trade-offs.

[21] Costs and risk allocation are also ways of regulating relations and conflicts between stakeholders.

Funding Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement. Novelli's work was supported by a grant provided by Fujitsu Limited to the Centre for Digital Ethics, Department of Legal Studies, Alma Mater – Università degli Studi di Bologna, rep. 95/2021, on "The Nature of Accountability in AI – Toward Responsible AI".

Data availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

Akpanuko EE, Asogwa IE (2013) Accountability: a synthesis. Int J Finance Account 2(3):164–173

Barclay I and Abramson W (2021) Identifying roles, requirements and responsibilities in trustworthy AI systems. In: Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers (pp 264–271). Association for Computing Machinery. https://doi.org/10.1145/3460418.3479344

Baxter G, Sommerville I (2011) Socio-technical systems: from design methods to systems engineering. Interact Comput 23(1):4–17. https://doi.org/10.1016/j.intcom.2010.07.003

Binns R (2018) Algorithmic accountability and public reason. Philos Technol 31(4):543–556. https://doi.org/10.1007/s13347-017-0263-5

Bovens M (2007) Analysing and assessing accountability: a conceptual framework. Eur Law J 13(4):447–468. https://doi.org/10.1111/j.1468-0386.2007.00378.x

Bovens M (2010) Two concepts of accountability: accountability as a virtue and as a mechanism. West Eur Polit 33(5):946–967. https://doi.org/10.1080/01402382.2010.486119

Busuioc M (2021) Accountable artificial intelligence: holding algorithms to account. Public Adm Rev 81(5):825–836. https://doi.org/10.1111/puar.13293

Cooper AF, Moss E, Laufer B and Nissenbaum H (2022) Accountability in an algorithmic society: relationality, responsibility, and robustness in machine learning. In: 2022 ACM Conference on Fairness, Accountability, and Transparency, 864–876. https://doi.org/10.1145/3531146.3533150

Desouza K, Dawson G, Chenok D (2019) Designing, developing, and deploying artificial intelligence systems: lessons from and for the public sector. Bus Horizons. https://doi.org/10.1016/j.bushor.2019.11.004

Dubnick M (2005) Accountability and the promise of performance: in search of the mechanisms. Public Perform Manage Rev 28(3):376–417

Durante M, Floridi L (2022) A legal principles-based framework for AI liability regulation. In: Mökander J, Ziosi M (eds) The 2021 yearbook of the digital ethics lab. Springer International Publishing, pp 93–112. https://doi.org/10.1007/978-3-031-09846-8_7

Ehsan U, Liao QV, Muller M, Riedl MO and Weisz JD (2021) Expanding explain ability: towards social transparency in AI systems. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp 1–19. https://doi.org/10.1145/3411764.3445188

Endsley MR, Kaber DB (1999) Level of automation effects on performance, situation awareness and workload in a dynamic control task. Ergonomics 42(3):462–492. https://doi.org/10.1080/001401399185595

Fjeld J, Achten N, Hilligoss H, Nagy A and Srikumar M (2020) Principled artificial intelligence: mapping consensus in ethical and rights-based approaches to principles for AI (SSRN Scholarly Paper Fasc. 3518482). https://doi.org/10.2139/ssrn.3518482

Floridi L (2013) Distributed morality in an information society. Sci Eng Ethics 19(3):727–743. https://doi.org/10.1007/s11948-012-9413-4

Floridi L (2016) Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. Philos Trans Roy Soc A 374(2083):20160112. https://doi.org/10.1098/rsta.2016.0112

Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M and Kagal L (2018) Explaining explanations: an overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 80–89. https://doi.org/10.1109/DSAA.2018.00018

Grant RW, Keohane RO (2005) Accountability and abuses of power in world politics. Am Polit Sci Rev 99(1):29–43. https://doi.org/10.1017/S0003055405051476

Hardin G (1968) The tragedy of the commons. Sci New Ser 162(3859):1243–1248

Kaler J (2002) Responsibility, accountability and governance. Bus Ethics 11(4):327–334. https://doi.org/10.1111/1467-8608.00292

Kaminski ME (2020) Understanding transparency in algorithmic accountability. In: Barfield W (ed) The Cambridge handbook of the law of algorithms. Cambridge University Press, pp 121–138

Kroll JA (2020) Accountability in computer systems. Oxford Handbook Ethics AI. https://doi.org/10.1093/oxfordhb/9780190067397.013.10

Lindberg SI (2013) Mapping accountability: core concept and subtypes. Int Rev Adm Sci 79(2):202–226. https://doi.org/10.1177/0020852313477761

Long S (2013) Socioanalytic methods: discovering the hidden in organisations and social systems. Routledge. https://www.routledge.com/Socioanalytic-Methods-Discovering-the-Hidden-in-Organisations-and-Social/Long/p/book/9781780491325

Martin K (2019) Ethical implications and accountability of algorithms. J Bus Ethics 160(4):835–850. https://doi.org/10.1007/s10551-018-3921-3

Mashaw JL (2006) Accountability and institutional design: some thoughts on the grammar of governance (SSRN Scholarly Paper ID 924879). Social Science Research Network. https://papers.ssrn.com/abstract=924879

Miller T (2019) Explanation in artificial intelligence: insights from the social sciences. Artif Intell 267:1–38. https://doi.org/10.1016/j.artint.2018.07.007

Mulgan R (2000) 'Accountability': an ever-expanding concept? Public Admin 78(3):555–573. https://doi.org/10.1111/1467-9299.00218

Mulgan R (2003) Issues of accountability. In: Mulgan R (ed) Holding power to account: accountability in modern democracies. Palgrave Macmillan, pp 1–35. https://doi.org/10.1057/9781403943835_1

Olsen JP (2017) Ambiguity and the politics of accountability. In: Olsen JP (ed) Democratic accountability, political order, and change: exploring accountability processes in an era of European transformation. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198800606.003.0004

Romzek BS, Dubnick MJ (1987) Accountability in the public sector: lessons from the challenger tragedy. Public Adm Rev 47(3):227–238. https://doi.org/10.2307/975901

Rubenstein J (2007) Accountability in an unequal world. J Polit 69(3):616–632. https://doi.org/10.1111/j.1468-2508.2007.00563.x

Schedler A (1999) Conceptualizing accountability. https://works.bepress.com/andreas_schedler/22/

Sinclair A (1995) The chameleon of accountability: forms and discourses. Acc Organ Soc 20(2):219–237. https://doi.org/10.1016/0361-3682(93)E0003-Y

Srinivasan R, San Miguel González B (2022) The role of empathy for artificial intelligence accountability. J Responsib Technol 9:100021. https://doi.org/10.1016/j.jrt.2021.100021

Taddeo M (2017) Trusting digital technologies correctly. Mind Mach 27(4):565–568. https://doi.org/10.1007/s11023-017-9450-5

Theodorou A, Dignum V (2020) Towards ethical and socio-legal governance in AI. Nat Mach Intell. https://doi.org/10.1038/s42256-019-0136-y. (**Art. 1**)

Thompson DF (1980) Moral responsibility of public officials: the problem of many hands. Am Polit Sci Rev 74(4):905–916. https://doi.org/10.2307/1954312

Thynne I, Goldring J (1987) Accountability and control: government officials and the exercise of power. Law Book Company

Tsamados A, Aggarwal N, Cowls J, Morley J, Roberts H, Taddeo M, Floridi L (2022) The ethics of algorithms: key problems and solutions. AI Soc 37(1):215–230. https://doi.org/10.1007/s00146-021-01154-8

Vermaas P, Kroes P, van de Poel I, Franssen M, Houkes W (2011) A philosophy of technology: from technical artefacts to sociotechnical systems. Synthesis Lectures Engineers Technol Soc 6(1):1–134. https://doi.org/10.2200/S00321ED1V01Y201012ETS014

Zech H (2021) Liability for AI: public policy considerations. ERA Forum 22(1):147–158. https://doi.org/10.1007/s12027-020-00648-0

Zweig KA, Raudonat F (2022) Accountability of artificial intelligence in human resources. In: Strohmeier S (ed) Handbook of Research on Artificial Intelligence in Human Resource Management. Edward Elgar Publishing, 323–336