Taming Artificial Intelligence: A theory of control-acountability alignment among AI developers and users

Gudela Grote

ETH Zürich

ggrote@ethz.ch

Sharon K. Parker

Curtin University

s.parker@curtin.edu.au

Kevin Crowston

Syracuse University

crowston@g.syr.edu

Acknowledgements

We thank Associate Editor Ruth Aguilera for her thoughtful guidance throughout the review process, and three anonymous reviewers for their constructive feedback, which greatly helped to bring this work to fruition. We also thank participants at the NSF Working in the Age of Intelligent Machines (WAIM) Network Conference, Washington DC, June 6 - 7, 2022, for the engaged discussions of ideas that are at the core of this paper. Kevin Crowston acknowledges support by the NSF grant "The Future of News Work: Human-Technology Collaboration of Journalistic Research and Narrative Discovery" (21-29047).

Accepted for publication in Academy of Management Review.

TAMING ARTIFICIAL INTELLIGENCE: A THEORY OF CONTROL–ACCOUNTABILITY ALIGNMENT AMONG AI DEVELOPERS AND USERS

ABSTRACT

The growing agency of artificial intelligence (AI) systems, more specifically systems based on machine learning, has raised concern about the security, safety, and ethical risks of AI use. We argue that core to mitigating AI risks is proper alignment of control and accountability for the stakeholders involved in AI development and use. Control enables and accountability motivates stakeholders to achieve desired and avoid undesired outcomes using AI. However, AI systems' capabilities for autonomous adaptivity reduce control even for the experts who create them. Moreover, increasing interdependencies between AI development and use render it difficult to unambiguously locate control and accountability. In this paper, we address these challenges for mitigating AI risks by postulating decentralized forms of stakeholder governance and integrative negotiations among stakeholders during the AI life cycle as conducive to aligning control and accountability for AI development and use. Further, we specify that extensive information sharing aided by perspective taking and a shared norm of accountability facilitate integrative negotiation strategies. We conclude by discussing the implications of our theory for management scholarship on the impact of AI and identify promising avenues for future research at micro, meso, and macro levels of analysis.

TAMING ARTIFICIAL INTELLIGENCE: A THEORY OF CONTROL–ACCOUNTABILITY ALIGNMENT AMONG AI DEVELOPERS AND USERS

With the advent of autonomous and adaptive artificial intelligence (AI) systems discussions about its impact abound. AI, defined as "systems that build on machine learning, computation, and statistical techniques, as well as rely on large data sets to generate responses, classifications, or dynamic predictions" (Faraj, Pachidi & Sayegh, 2018, p. 62), has sparked imaginations in unparalleled ways due to dueling visions of AI as the new steam engine, creating boundless possibilities for businesses and society to prosper, or as "a leviathan that must be restrained and deployed with extreme caution in order to prevent it from taking over and killing us all" (Roose, 2023). Governments and regulatory bodies have started to develop requirements for managing the risks of AI (e.g., European Union, 2023; NIST, 2023; The White House, 2023) and high-level expert groups meet regularly to foster international agreements on measures to alleviate AI risks (Hern, 2024). However, the speed of technological innovation complicates such efforts. Public availability of ChatGPT and similar large language models (LLMs) has added fervor to the discussions of AI-related risks, culminating in a call for a moratorium on further development to allow technology developers, policy makers, and regulatory bodies to devise measures to contain the risks of such systems (Future of Life Institute, 2023).

Current debates in the management literature mirror the two opposing visions of AI. Some authors fear profound negative consequences for employees, organizations, and society due to, for example, increased surveillance, disinformation, rising inequality, and curtailed human agency (Balasubramanian, Ye, & Xu, 2022; Jarvenpaa & Välikangas, 2020; Kane, Young, Majchrzak, & Ransbotham, 2021; Kellogg, Valentine, & Christin, 2020; Zuboff, 2019). Others offer a more optimistic outlook, especially for firms' value creation (Brock & von Wangenheim, 2019; Murray, Rhymer, & Sirmon, 2021; Shresta, Ben-Menahem, & von

Krogh, 2019), but also for society at large, such as using AI to improve healthcare and fight climate change (Chhillar & Aguilera, 2022; Floridi, Cowls, King, & Taddeo, 2021).

It is received wisdom in socio-technical design thinking (e.g., Clegg, 2000; Hollnagel & Woods, 2005; Leonardi, 2012) that the effects of technology depend on a multitude of factors, from specific features of the technology itself to how it is embedded and made sense of by users in their daily work (e.g., Bailey & Barley, 2020; Leonardi & Barley, 2010; Orlikowski & Scott, 2008; Parker & Grote, 2022). The concept of human-in-the-loop, postulating that humans must have control over a technical system to fulfill their accountability for its safe and effective performance, is considered key for creating positive impact through technology (Billings, 1997; Endsley, 2023; Hollnagel & Woods, 2005). However, providing humans with sufficient control is easier said than done, as evidenced by many failures of complex technologies, such as the tragic crashes of the Boeing 737 Max when pilots were unable to override actions taken by flawed software. The Boeing story also illustrates how decisions on technological design are influenced by business interests that can supersede the concern for adequate human-technology interaction (Norman & Euchner, 2023).

In the case of AI, the same concepts and challenges for control and accountability have surfaced. Human control is argued to be fostered by explainable and interpretable AI that should be deployed to augment rather than replace human decision-making (e.g., Choudhary, Marchetto, Shrestha, & Puranam, 2023; Langer et al., 2021; Raisch & Fomina, 2023; Rudin, 2019). Questions of accountability, paired with fundamental ethical concerns, have received unprecedented attention, fuelled especially by the emerging impact of AI on the broader society beyond workers (Floridi et al., 2018). AI is also a textbook example of how business interests drive technology development (Roose, 2023).

However, there are some important technical and organizational specificities that warrant reconsideration of how control and accountability can be aligned for AI-based technologies, thus motivating our theorizing. The most crucial difference is that contemporary

AI development involves learning from large data sets produced before *and* during the use of the new AI system, making AI the first technology that is fundamentally reshaped by its use (e.g., Faraj et al., 2018; Jacobides, Brusoni, & Candelon, 2021; Slota et al., 2023). The complexity and dynamic nature of self-learning AI systems drastically reduces their transparency and predictability, creating the so-called 'black-box problem' and reduced control even for AI developers (Asatiani, Malo, Nagbol, Penttinen, Rinta-Kahila, & Salovaara, 2021; Berente, Gu, Recker, & Santhanam, 2021; Castelvecchi, 2016; Diakopoulos, 2016; Rudin, 2019). For instance, after Google reportedly fixed the problem of its image-recognition algorithm classifying black people as gorillas by simply blocking this image category, it remained unclear whether software developers had chosen this quick-fix because they were not able to find a more complete solution, because they did not allocate sufficient resources to do so, or for some other reason (Vincent, 2018). Similarly, LLM developers' problems with adding 'guardrails' to their systems or to eliminate hallucinations, that is irrelevant, incorrect, or made-up information, have led researchers to call for AI-resilient interfaces that allow better contextualization of LLM outputs (Glassman, Gu, & Kummerfeld, 2024).

Moreover, the tight relationship between development and use blurs the lines between what AI developers and AI users should be able to control and held accountable for (Wieringa, 2020). Managing biases in training data is one frequently discussed example for how developers and users must rely on each other to ensure valid algorithms and system outcomes (e.g., Choudhary et al., 2023; Teodorescu, Morse, Awwad, & Kane, 2021). However, there is frequently a whole AI supply chain to be considered, involving data providers, companies offering access to large AI models through cloud-based services, or knowledge brokers helping to translate system outputs for AI users, which creates a "many hands" problem for allocating accountability (Cobbe, Veale, & Singh, 2023; Waardenburg & Huysman, 2022; Waardenburg, Huysman, & Sergeeva, 2022).

With the growing capabilities of AI systems for autonomous adaptivity, it appears that we are approaching a situation where no human has full system control anymore. Yet, by today's legal standards, and for the foreseeable future it must be humans who are held to account for outcomes of AI systems (Burton, Habli, Lawton, McDermid, Morgan, & Porter, 2020). Accordingly, reduced control for all stakeholders involved in AI development and use raises new challenges for how accountability is allocated. Also, increasing interdependencies between stakeholders renders it more challenging to unambiguously locate control and accountability. In what follows, we argue, consistent with existing socio-technical principles, that an alignment between control and accountability will reduce AI risks, but how that alignment is achieved and facilitated needs re-thinking. Thus, we ask: how can control and accountability be aligned for AI systems to mitigate AI risks? It is this fundamental question we address in our theory.

To answer this question, we integrate research on control and accountability, stakeholder governance, and stakeholder negotiations. We outline how control-accountability alignment helps to mitigate AI risks because control enables, and accountability motivates, stakeholders to achieve desired and avoid undesired outcomes through their actions. AI risks concern potential negative consequences of employing AI, such as biased personnel selection, faulty medical diagnoses, or unfair treatment of employees or customers, but also risks stemming from misaligned control and accountability itself, such as stress and reduced well-being for the involved workers and financial or reputational losses for the involved organizations. We discuss the consequences of different degrees of AI systems' autonomous adaptivity for shifts in AI users' and developers' control and accountability and specify conditions that make effective decisions on aligning control and accountability, and thereby adequate risk mitigation, more likely. In particular, we propose more decentralized forms of stakeholder governance and integrative negotiations that engage stakeholders during the full AI life cycle

as prerequisites for achieving control-accountability alignment during AI development and use.

Our theory contributes to management scholarship on the impact of AI in several ways. First, at the micro-level of AI workers' activities, we offer a deeper understanding of how AI developers and users can together mitigate AI risks. To date, concerns about control-accountability alignment have almost exclusively focused on the working conditions for users of AI technologies (Parker & Grote, 2022). We argue that increased task interdependencies with AI make this perspective overly narrow, and that significant risks emerge if control-accountability issues are not considered for AI developers as well. Second, at the meso-level of organizational functioning, we draw on the organizational control and governance literatures (Aguilera, Filatotchev, Gospel, & Jackson, 2008; Sitkin, Long, & Cardinal, 2020) to detail how control and accountability for advanced AI systems can be aligned across different actors in different organizations and suggest organizational measures in support of such an alignment. We thus offer ways to manage the "many hands" problem in the AI supply chain (Cobbe et al., 2023). Third, at the macro-level of stakeholder relationships, we propose conditions for integrative stakeholder negotiations within more decentralized forms of governance. We thereby add to recent debates about effective ways for handling power differentials among stakeholders in AI governance (Chhillar & Aguilera, 2022).

Overall, our theory aligns with recent calls to address the development and use of technology in a more integral fashion, including issues of power, diverging perspectives on the purpose and effects of technology, and limited predictability of emerging use patterns (Anthony, Bechky, & Fayard, 2023; Bailey & Barley, 2020). As a caveat, we do not consider the kind of use for which AI is intended: we rather specify conditions that make it more likely that the intended use can be successfully realized. Accordingly, our theory does not directly contribute to current debates about risks for AI being employed for malevolent purposes, for instance in relation to algorithmic management (Kellogg et al., 2020) or undue surveillance

(Zuboff, 2019). However, allocating accountability for achieving desired and avoiding undesired outcomes of AI use to AI developers in cases where AI users have no or little control, may make development of malevolent AI less likely.

CONTROL AND ACCOUNTABILITY IN ORGANIZATIONS

Control and accountability are ubiquitous concepts in a wide range of literatures. In psychological and organizational research, control is usually defined in terms of an actor being enabled to achieve desired outcomes, often with reference to self-regulatory processes of goal striving (Brehmer, 1992; Carver & Scheier, 1990; Green & Welsh, 1988; Skinner, 1996). Especially in psychological models of control, it is emphasized that control involves influence over a current situation, but also a sufficient understanding of how to proceed in the situation, arising from transparency of ongoing processes and predictability of future states and outcomes (Hollnagel & Woods, 2006; Skinner, 1996). Thus, from this perspective, individuals have control in a situation when they understand what is happening and can influence that situation in predictable ways.

A distinct perspective on control at the organizational level discusses how achieving desired outcomes entails mechanisms to align individuals' and groups' behaviors with the goals of the organization. Thereby, a second meaning of control as constraining the actions of others is invoked, often referred to as managerial or organizational control (Merchant & Otley, 2006; Ouchi, 1979; Sitkin et al., 2020). Such control usually spans hierarchical levels within organizations and reflects differences in power, defined as asymmetric control over valued resources in a social relationship (Magee & Galinsky, 2008). It is the former perspective of control, as influence, transparency, and predictability, that we emphasize when we discuss control over AI systems.

Accountability on the other hand has been defined as "a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences"

(Bovens, 2007: 447). Put simply, accountability is about who is held to account for decisions made or actions taken, especially when things go wrong. The forum to which an actor must answer can vary: it may be their employer, their peers, a regulatory agency, the courts, or even civil society at large. Accountability is assumed to motivate actors to use their control in line with objectives and rules of the social system within which they act (Bovens, 2007; Frink et al., 2008; Hall, Frink, & Buckle, 2017). It is thus directly linked to managerial or organizational control because making individuals answerable for their behavior is a key mechanism for aligning behavior with organizational objectives (Frink et al., 2008). Agency theory is very explicit about this relationship by postulating that agents, such as managers in an organization, may be motivated more by self-interest than by the interests of their principal, such as a firm's shareholders (Eisenhardt, 1989). Accountability, along with rewards and punishments for (not) furthering the principal's objectives, is assumed to incentivize agents to align their actions with superordinate goals.

Accountability also operates at higher levels of governance where organizations are held to account to foster alignment with objectives by internal and external stakeholders and society at large. "In this context [of corporate governance], effectiveness in the broadest sense involves the accountability of corporate decision-makers and the legitimacy of decisions about their different economic and noneconomic goals and values" (Aguilera et al., 2008: 476). Firms' striving for control to gain competitive advantage is bounded by legal accountability and liability, but also by the responsibility for joint creation of both economic and social value for stakeholders (Bacq & Aguilera, 2022; Bridoux & Stoelhorst, 2022; McGahan, 2023).

The psychological and organizational accountability literatures rarely consider whether the entity being held accountable has indeed sufficient control to achieve the desired and avoid the undesired outcomes. Accordingly, the principle that actors should only be required to explain and justify actions they themselves understand and have an influence over is not

frequently addressed in related research (Merchant & Otley, 2006). In the few studies that have addressed the potential misalignment of holding actors accountable for things they have no control over, researchers have found that the stress-inducing effect of accountability is buffered by higher levels of job autonomy, that is control over one's own work (Hall, Royle, Brymer, Perrewé, Ferris, & Hochwarter, 2006). In a related vein, Lerner and Tetlock (1999) discuss findings that cognitive biases in decision-making are only reduced by felt accountability if the decision-makers understand these biases and their effects. Moreover, it has been argued that accountability combined with control increases the willingness to accept risk rather than to transfer risk to others (Eisenhardt, 1989). Frink et al. (2008) have developed an elaborate model of how individual and organizational accountability can foster reputation, performance, and well-being if accountability requirements are matched with the appropriate resources and capabilities. In the most general terms, aligning control and accountability is key to achieving desired and avoiding undesired outcomes, thereby contributing to effective management of risks in organizations (Merchant & Otley, 2006).

CONTROL AND ACCOUNTABILITY IN AI DEVELOPMENT AND USE

From the discussion of control and accountability in organizations, we take three fundamental considerations as relevant for mitigating risks in the development and use of AI. First, risks are mitigated if actors are only held accountable for actions and outcomes that they have control over, or stated differently, accountability requires that actors are *enabled* by adequate means of control to achieve desired and avoid undesired outcomes. Second, accountability channels actors' use of their control towards actions and outcomes that are valued within the larger social system to which they belong, that is accountability *motivates* actors to act in accordance with superordinate interests. Third, misalignment can entail that an actor is accountable but has no or little control or that an actor has control but is not accountable. These two forms of misalignment can be interlinked if an actor in control transfers accountability to an actor with no control. In all cases of misalignment—as further

discussed also in relation to our second proposition—the risk of actors producing undesired outcomes are increased. Those who would be motivated to mitigate risks do not have sufficient control to do so, while those who are in control may be motivated by self-interest alone, which also leads to suboptimal outcomes overall and increased risk. These considerations are captured in our first proposition, which concerns the focal outcome of AI risk mitigation within our conceptual model as shown in Figure 1:

Proposition 1. To the extent that control and accountability during AI development and use are aligned, AI users and developers are enabled and motivated to achieve desired and avoid undesired outcomes and thereby mitigate AI risks.

Insert Figure 1 about here

Current debates about AI risks center on the potential for severe misalignments between control and accountability brought about by the autonomous adaptivity of AI systems. The complexity and dynamism of these systems reduces transparency, predictability, and influence for both human users and developers, which raises the fundamental concern of who can still effectively manage AI risks. Correspondingly, one needs to ask who can be held accountable for the correct functioning of AI systems and desired outcomes of their use. Answers to these questions in extant research vary, depending on what the actual or imagined capabilities of AI systems are.

Risks are considered manageable if AI is used to augment human users' capabilities rather than replace them and if transparency and predictability for the human users are ensured (Baird & Maruping, 2021; Choudhary et al., 2023; Crowston & Bollici, 2020; Murray et al., 2020; Raisch & Krakowski, 2021; Shresta et al., 2019). In such situations, control of and accountability for system outcomes rests with the AI user, what has been termed "human-in-the-loop" in human factors engineering (Billings, 1997; Endsley, 2017, Holford, 2022;

Hollnagel & Woods, 2005). Conditions for control by human users are most easily met if AI systems use simple algorithms, either programmed or trained with only few parameters, and if functions necessary for safe and effective task performance are allocated between humans and AI in a complementary fashion, that is in accordance with strengths and weaknesses of both (Challenger, Clegg, & Shepherd, 2013; Grote, Ryser, Wäfler, Windischer, & Weik, 2000; Hollnagel & Woods, 2005). The employed algorithms can be made sufficiently transparent and predictable for AI users through methods for explainable AI (Kim & Doshi-Velez, 2021; Mittelstadt, Russell, & Wachter, 2019; Rudin, 2019) so that they are aware of what data are used to train the system or what models and weighing factors underly the system's recommendations. AI users are also given means to influence the system, for instance by choosing algorithms for particular decisions or correcting the system in case of error. Additionally, organizational practices can permit and give time to AI users to complement AI-based recommendations with their own intuition or further data, and they can ensure that AI users are not sanctioned when overriding recommendations of the AI system and are encouraged to call for system adjustments if problems emerge (Crowston & Bolici, 2020; Strich, Meyer, & Fiedler, 2021).

Technical decisions on how to render the system controllable for AI users are in the hands of AI developers as part of the control they have over how the AI system is built. AI developers thus carry the accountability for system functioning including the provision of all necessary technical conditions for control over system outcomes by AI users, such as explainability. The described ways of ensuring control and accountability for both AI developers and users can be strengthened by senior management explicitly granting agency to them. Thereby, users can also be encouraged to push for and implement AI technologies they consider crucial for their work, such as in the case described by Hartmann and Beane (2024) of police investigators self-initiating the use of digital trace data analytics. All of these

considerations are captured in Table 1 in the column describing the distribution of control and accountability for AI systems with comparatively low autonomous adaptivity.

Critics of AI tend to discuss AI systems with higher capabilities for autonomous adaptivity and the ensuing challenges of risks being out of human control and accountability being unaccounted for (e.g., Balasubramanian et al., 2022; Jarvenpaa & Välikangas, 2020; Kane, et al., 2021; Lindebaum, Vesa, & Den Hond, 2020). Such systems come in different varieties that either render control more difficult for AI users only, or for both AI users and developers (Asatiani et al., 2021). If an AI system employs more complex and dynamic algorithms that preclude efforts for making them explainable to AI users, AI users have little or no control over system outcomes anymore, and accordingly cannot be held accountable. Accountability for system use thus shifts to AI developers who can assert some control over system outcomes by setting boundary conditions for system use, for instance by 'freezing' the system when a desired level of accuracy has been achieved. By interacting with a more stable and predictable system rather than one that is continuously evolving as it learns from new data, the likelihood of avoiding undesired and achieving desired outcomes is increased (Babic, Gerke, Eveniou, & Cohen, 2019).

AI developers may also employ algorithms and methods for training systems that avoid the black-box problem, for instance by using supervised learning or even reverting to rule-based algorithms. For instance, Jain (2023) found that the development teams that produced software for the Amazon Social Chatbot Alexa to lead 'engaging conversations' with humans mostly used highly scripted rule-based algorithms rather than LLMs. By doing so, they tuned the capacities of AI to a level that granted them sufficient control over system development, but also over system outcomes, that is, conversations staying within socially acceptable limits. This example shows that, by feeling accountable for outcomes of system use, which in the case of the Amazon Alexa Challenge was induced by stating the aim of creating engaging conversations, AI developers can be motivated to effectively use their control over system

development to also control system outcomes, thereby mitigating risks. AI developers' efforts to maintain control in accordance with their accountability can be bolstered if senior management explicitly imparts decision-making power to them and establishes organizational mechanisms that prevent AI users from abusing whatever little control they may have left over system outcomes. Such mechanisms can be installed, for instance, as part of performance management by specifying for which purposes and objectives an AI system should be used (Hall, Frink, Ferris, Hochwarter, Kacmar, & Bowen, 2003). These considerations are captured in Table 1 in the column describing the distribution of control and accountability for AI systems with medium levels of autonomous adaptivity.

Especially when using deep neural networks, not only AI users', but also AI developers' control can be severely curtailed due to the inability to understand and predict the behavior of these highly complex and autonomously learning algorithms. If accountability still rests with AI developers in such cases, they are motivated to use their remaining influence to shape the AI system in ways that compensate for their limited control. Asatiani et al. (2021) have proposed the use of operating envelopes for this purpose. For instance, a face recognition system might be restricted to make decisions only for photos taken with good lighting conditions if the training data for the system included only such photos. Also, certain undesirable outputs (e.g., speeding in a self-driving car) can be excluded. Users of these highly complex and opaque systems usually will have no control anymore and have to fully rely on the system as part of their work tasks. However, they may contribute with their daily experience to the intense monitoring and testing by AI developers as part of their efforts to ascertain safe and effective system performance.

In the extreme, AI developers may see themselves confronted with their senior management's desire to go to the limits of what may be possible with AI to gain competitive advantage, with no adequate safeguards for testing and monitoring such systems. A key safety researcher leaving OpenAI just days after the launch of their most recent AI model GPT-4o is

a case in point (Milmo, 2024). If AI developers are thus required to work at the very edge of their own capabilities, severely curtailing their control over the systems they develop, accountability for system functioning and outcomes rests with senior management. By holding senior management accountable, their motivation for establishing sound processes for risk mitigation and strong ties to independent oversight bodies is likely to increase (Shneiderman, 2016). Similarly, they will be more ready to set up internal mechanisms to vitalize all resources in the organization for coping with everyone's limited control capabilities and to foster continuous learning. Such internal mechanisms have been proposed for "managing the unexpected" more generally and can be applied to mitigate AI risks as well: proactively search for aberrations in operations; build capacity to bounce back and learn from failure; and welcome expertise from anywhere in the organization (Weick & Sutcliffe, 2001). These considerations are captured in Table 1 in the column describing the distribution of control and accountability for AI systems with the highest level of autonomous adaptivity.

In summary, control-accountability alignment for AI development and use requires an expanded socio-technical perspective that considers the technology not in isolation or with a narrow focus on the interaction between the technology and its users, but as part of a broader system of actors that influence the governance of technology development and use (Anthony et al., 2023; Asatiani et al., 2021). Adopting this expanded perspective early and proactively helps to raise awareness of the linkages between the technology and opportunities and constraints for control-accountability alignment as outlined in Table 1. Even in the presumably simplest case, where AI systems are still controllable by human users, control-accountability alignment requires close attention to ensure that AI developers provide AI users with all necessary means for understanding and influencing the system. Being held accountable for both system functionality and system outcomes in the cases of more autonomous and adaptive AI systems fosters AI developers' motivation to handle their own and the users' limitations of control responsibly and to introduce the necessary measures to

mitigate the increased risks of these systems, such as monitoring and feedback systems (Bartsch, Milani, Adam, & Benlian, 2024) and definition of operating envelopes (Asatiani et al., 2021). Senior management of user and developer organizations are accountable for empowering individual AI users and developers and to put systems in place for continuous learning, such as "red teaming", where users emulate real attackers' techniques, or "sandboxes" in which systems can be tested in a safe environment. Another, rather special example of creating feedback loops for continuous learning are the ongoing efforts of both lay people and computer scientists to test the reliability and veracity of LLMs' outputs (Srivastava et al., 2023). In other words, the solution for a potentially out-of-control technical system is to embed it in a more sophisticated organizational system of control and accountability. This discussion leads to our second proposition:

Proposition 2. To the extent that AI systems' capabilities for autonomous adaptivity are proactively and jointly considered alongside challenges for AI users' and developers' control and accountability, control-accountability alignment is more likely and AI risks are mitigated.

Insert Table 1 about here

STAKEHOLDER ENGAGEMENT FOR CONTROL-ACCOUNTABILITY ALIGNMENT

Having theorized the relevance of control-accountability alignment for mitigating AI risks, or *why* this topic matters, and having proposed *what* key considerations are required for effective decision-making on alignment, we now turn to *how* AI stakeholders may indeed make such decisions (see Figure 1 for the complete conceptual model). In current discussions of AI risk management, regulators and other independent oversight bodies are considered key stakeholders, as companies developing AI are not trusted to undertake the necessary steps for

mitigating risk of their own accord (Falco et al., 2021; Shneiderman, 2016). At the same time, these discussions usually acknowledge that effective regulation for fast-moving technologies is difficult to achieve, therefore requiring immediate action by stakeholders directly involved in AI development and use in any case (Bengio et al., 2024; Chhillar & Aguilera, 2022).

In our theorizing, drawing on this latter argument, we focus on the stakeholders most closely involved in AI development and use, such as developers employed in one organization who develop and implement an AI system for users employed in another organization. Even such a simple case may already span a large group of people and several organizations, such as providers of training data, model developers and testers, user interface designers, professional users of AI models, and users working with outputs from AI models without directly interacting with the models. Besides the individuals and teams involved in the operative work processes of AI development and use, the senior management of their organizations also have stakes in what and how AI is developed and used. Together, these stakeholders have the most knowledge of the technology and the context within which it is being implemented and can therefore provide the most relevant inputs into decision-making, while also being directly impacted by the decisions taken. Individuals who are the targets of AI-based outcomes, such as employees being (not) hired or bank clients being (not) given a loan, are another crucial stakeholder group, whose interests we indirectly consider by positing that control-accountability alignment is conducive to achieving desired and avoiding undesired outcomes. Overall, our focus is on deriving conditions that make it more likely that the most immediately involved stakeholders will engage in effective risk mitigation, in particular via decentralized forms of stakeholder governance and integrative negotiation strategies for decisions on control-accountability alignment.

Decentralized governance

As Slota et al. (2023: 1287) have argued, "socio-technical landscapes" rather than bounded organizations need to be considered for adequate management of AI risks. We

cannot capture the full complexity for governing AI implied by this broadened perspective which entails an intricate interplay of technical, economic, legal, and societal constraints and requires consideration of the varied organizational settings in which AI development and use may take place (Chhillar & Aguilera, 2022). We argue more modestly that decisions on control-accountability alignment need to be curated with a constant awareness of the socio-technical landscape within which they are situated. The overarching purpose is to bring the necessary knowledge and the diversity of perspectives and interests to bear on decisions concerning AI capabilities and matching allocations of control and accountability among AI developers and users over the full AI life cycle.

When stakeholders decide on the allocation of control and accountability for AI development and use, they are confronted with the unique characteristic of AI systems evolving through their use: "Data collection is ongoing, and the landscape of data is ever-shifting and rarely ideal (...) the affordances of the system can continually produce novel, counter-intuitive, and unpredictable uses once deployed" (Slota et al., 2023: 1292). The entanglement of AI development and use raises challenges because it increases the interdependencies between the tasks of AI developers and users and makes it harder to unequivocally locate accountabilities and the control needed to answer to them (Cobbe et al., 2024; Holford, 2022; Raisch & Krakowski, 2020; Waardenberg & Huysman, 2022). For instance, AI developers may be held accountable for minimizing biases in the systems they create, but they must often rely on training data scraped from the Internet or created in other ways they cannot control (Chan, Bradley, & Rajkumar, 2023). Or they may have to explain to users, which algorithms are employed in their systems and with what effects, while at the same time these algorithms adapt to the ways users employ the systems (Dolata & Crowston, 2023). These examples illustrate how control and accountability may be distributed across actors due to only partial control by all.

Recent literature on stakeholder governance addresses exactly this issue of managing control and accountability across actors for complex and dynamic processes of joint value creation that AI development and use is a prime example of (Amis, Barney, Mahoney, & Wang, 2020; Bacq & Aguilera, 2021; Bridoux & Stoelhorst, 2022). Governance concerns the sets of rules for how decision-making authority, responsibility for monitoring and sanctioning of rule violations, and lastly the jointly created value are distributed among stakeholders (Bridoux & Stoelhorst, 2022). In the most general terms, these rules specify how control and accountability for governance-related activities are allocated (Bacq & Aguilera, 2021). In centralized forms of governance, control and accountability rests with a focal firm or more precisely with the senior management of that firm, who set the rules, control and sanction rule compliance, and handle conflicts with individual stakeholders. Relationships among stakeholders are mostly dyadic between the focal firm and each stakeholder. In decentralized governance, all stakeholders interact with each other based on commonly agreed upon rules and mechanisms for rule compliance and shared control and accountability, which requires that they develop trust in the system of rules they have created rather than in the managers of a focal firm. As an example, Chen, Richter and Patel (2021) discuss the governance of digital platforms and find that application platforms, such as Uber or Facebook, are more likely to be governed centrally by platform owners, whereas infrastructure platforms, such as the Internet or block-chain infrastructures, tend to be governed through collective efforts by platform participants, that is end users and third-party developers. They propose that the push for more decentralized governance of infrastructure platforms stems from the needs for unhindered collaborative innovation and for protecting such infrastructures as an important resource for many.

We follow a similar argument to that of Bridoux and Stoelhorst (2022) stating that decentralized forms of governance are best suited for joint value creation activities that involve high levels of complexity and environmental dynamism. If stakeholders are jointly in

control and accountable for all decisions taken along the AI life cycle, it is more likely that they will share the fundamental concern for how AI risks are mitigated. This argument aligns with Bac and Aguilera's (2021) emphasis of deliberation as a key process for stakeholder governance that permits integration of legitimate stakeholder interests based on different mechanisms for handling stakeholder power, allowing to empower stakeholders (e.g., individual AI users), to curtail power (e.g., organizations owning data and computing resources), or endorse power of additional stakeholders (e.g., regulators). Free-rider problems, such as when some stakeholders have control but are not held accountable, and exploitation of less powerful actors, such as by holding them accountable even though they have little or no control, can be overcome if actors are willing to collaborate based on a common understanding that actions guided by pure self-interest will result in suboptimal results for everyone (Bridoux & Stoelhorst, 2022). This discussion leads to our third proposition:

Proposition 3. To the extent that stakeholders enact more decentralized forms of governance to reach decisions on allocating control and accountability during AI development and use, control-accountability alignment is more likely and AI risks are mitigated.

Integrative negotiations

Stakeholders likely will differ substantially regarding their objectives and perspectives on any AI system to be developed along with desirable allocations of control and accountability as well as regarding the power that they can exert to push these objectives and perspectives. Decisions therefore must be reached through negotiations, that is a "process for resolving a wide variety of disagreements over both tangible and intangible interests among two or more parties with common interests to motivate finding a mutually acceptable solution" (Churchman, 2019). Similarly, proponents of newer more decentralized forms of governance have stressed deliberation and bargaining to reach decisions (Amis et al., 2020; Bacq & Aguilera, 2021; Bridoux & Stoelhorst, 2022).

In these negotiations, stakeholders are not only actors to be held to account but may also represent a forum to which other actors have to answer and be accountable. Van den Broek, Levina, and Sergeeva (2022) have provided an intricate account of how control and accountability were continuously renegotiated between AI developers and users during the development of an AI-based hiring tool. They describe the tensions that arose when managers and employees were requested to deliver data on their performance assessments to data scientists who had been mandated to develop the tool. The HR team acted as a boundary spanner in its attempts to alleviate these tensions by convincing managers and employees of the higher quality of data-driven hiring decisions, but also by explaining to data scientists the legal and ethical constraints for data access. The HR team established a new company-wide standard for performance evaluations to be able to deliver adequate data to the data scientists. This example illustrates the multitude of decisions that are part of aligning control and accountability for all stakeholders as prerequisite for the effective employment of AI.

The main objective for curating control-accountability negotiations is to bring stakeholders to adopt an integrative negotiation strategy that focuses on producing joint gains, as opposed to a distributive negotiation strategy based on fixed-pie perceptions of one party's gain being another's loss. In order to avoid that everyone tries to maximize their control and minimize their accountability, which would happen in a distributive negotiation strategy, stakeholders are encouraged to develop what Curhan, Overbeck, Cho, Zhang, and Yang (2023) have termed a deliberative mindset. They become aware of interdependencies in their tasks and the resulting entanglement of control and accountability and realize that giving up certain control to other stakeholders may help them to respond better to their own accountabilities. An integrative negotiation strategy may raise awareness in AI developers and their senior management that by making their systems more transparent and predictable, the likelihood of liability claims is reduced because everyone can work with the system more effectively with fewer errors. Such considerations may have contributed to sixteen big tech

companies recently pledging at a global AI summit to publish safety frameworks for measuring AI risks (Hern, 2024).

The scope and complexity of a new system, and different options for control-accountability alignment, are negotiated for the first time when the problem the AI is to solve is defined and the business opportunities and risks associated with the AI are elaborated. These negotiations continue throughout the AI life cycle as experience gained during system development and use leads to adaptations in the system and in the allocation of control and accountability. Decisions are made and possibly revisited as to whether AI users should be given control and also held to account for system outcomes (Van den Broek, Sergeeva & Huysman, 2021).

A key challenge is to avoid "ironies of automation" (Bainbridge, 1983)—that is humans becoming the last resort for averting failures in technical systems that were designed to outperform them. The history of aviation is marred with examples of this fundamental challenge, the crash of Air France Flight 447 being a recent one. In this case, the pitot tubes of an Airbus 330 iced over in a thunderstorm, inducing the flight computers to transition into a mode that required manually flying, which overwhelmed the pilots' ability to comprehend the situation and therefore to act correctly, resulting in a crash (Holford, 2022; Oliver, Calvard, & Potocnik, 2017). In the realm of AI-based systems, similar concerns have been voiced, for instance regarding self-driving cars (Elish, 2019; Endsley, 2023). Another example is physicians who must base diagnostic and treatment decisions on poorly understood AI systems, while still living up to their accountability towards patients and fellow physicians (Lebovitz, Lifshitz-Assaf, & Levina, 2022).

Decisions on control-accountability alignment may follow from choices for more or less capable AI, but the reverse is also possible: decisions on the capabilities of an AI system may be made based on the desired control-accountability alignment. For instance, regulatory requirements may demand that users be in control and be held accountable, as is currently the

case in the medical field (Habli, Lawton, & Porter, 2020; Lebovitz, Levina, & Lifshitz-Assaf (2021). Moreover, firms may find that they can gain competitive advantage with less complex AI technology because systems can be more easily tailored to specific firm capabilities and objectives (Kemp, 2023).

Besides AI capabilities and the allocation of control and accountability to AI users and/or developers, a third important component of stakeholder negotiations concerns supporting mechanisms at the organizational level (see Table 1). Again, joint gains for all stakeholders can be realized, for instance by pairing AI developer accountability with performance management systems for AI users that prevent them from misusing their remaining, albeit very restricted, control in more advanced AI systems (Hall et al., 2003). This discussion leads to our fourth proposition:

Proposition 4. To the extent that stakeholders follow an integrative negotiation strategy as compared to a distributive strategy in decisions on the allocation of control and accountability, control-accountability alignment is more likely and AI risks are mitigated.

The negotiation literature mentions a wide range of factors that make integrative strategies more likely (Thompson, Wang, & Gunia, 2010). Two factors have been identified as particularly important: extensive information sharing helped by perspective taking and effectively managing power differentials between stakeholders. Regarding the first factor, information sharing has been shown to support integrative negotiation strategies because having a fuller picture of different stakeholders' goals and priorities helps to reduce fixed pie perceptions (Brett & Thompson, 2016; De Dreu, Beersma, Stroebe, & Euwema, 2006; De Dreu, Koole, & Steinel, 2000). In negotiations about control and accountability, it is essential that stakeholders know as much as possible about each other's intentions and expectations regarding the AI technology in question and openly discuss perceived conflicts between different objectives and ways to overcome them.

Information sharing is fostered by deliberately introducing prospective, real-time, and retrospective considerations as has been stressed in the risk literature (Hardy, Maguire, Power, &Tsoukas, 2020). If risk assessments are made prospectively, using tools for qualitative and quantitative risk modeling (NIST, 2023; Taddeo & Floridi, 2018), real-time experience with these systems is gained early within controlled settings, as in sandbox exercises (Gasser, 2024), and system failures are scrutinized retrospectively through incident reporting or root cause analysis (Macrae, 2022), all stakeholders develop a much more fine-grained understanding of what it takes to manage AI risks well and to align control and accountability accordingly. In the case of highly autonomous and adaptive AI systems, tight feedback loops between developers and users across the full AI life cycle can help to proactively identify problems before they lead to undesired outcomes and to develop a shared understanding of the system's objectives and limitations (Van den Broek, Sergeeva, & Huysman, 2021). Such a shared understanding helps decision-making on allocating control and accountability in an integrative manner because everyone becomes aware of the opportunities and constraints for effective risk mitigation by different actors involved in developing and using the system.

Examining AI risks in such detail is also promoted by stakeholders' willingness to take others' perspectives (Galinsky, Maddux, Gilin, & White, 2008). When people actively try to see the world from the view of another, several consequences emerge that foster more effective negotiations, including greater willingness to disclose information, improved trust, enhanced interpersonal problem-solving, lowered chance of conflict, and the propensity to engage in a win-win focused negotiating style (Parker, Atkins, & Axtell, 2008). In the risk literature, perspective taking has been stressed for containing risks in collaborative efforts which Weick and Roberts (1993) have termed heedful interrelating. This type of interaction is characterized by actors constantly (re)considering the effects of their own actions on the goals and actions of others.

For technology development and use, it has been argued that perspective taking is imperative, especially for bridging rationalist approaches to technology that focus on scientific knowledge, objectivity, and quantification with constructivist views that emphasize subjective meaning making in social discourse (Anthony et al., 2023; Jasperson, Carte, Saunders, Butler, Croes, & Zheng, 2002; Leonardi & Barley, 2010). Rationalist perspectives stress deterministic influences of technology on organizational and work processes and reduce organizations to production systems for enhancing efficiency and adaptability, in which accountability is assigned based on instrumental motives (Makarius, Mukherjee, Fox, & Fox, 2020; Murray et al., 2021). Constructivist perspectives highlight the entanglement of technology and social reality, as well as the emergent nature of new practices and routines. Based on value-oriented reasoning, accountability becomes a contested entity in continuous processes of adaptation that are difficult to predict and proactively shape (Elish, 2019; Leonardi & Barley, 2010; Orlikowski & Scott, 2008; Suchman, 2002). Acknowledging both rationalist and constructivist perspectives as equally relevant and valid helps to build a more complete understanding of how the technological impact on the organization of work can and should be shaped. This understanding constitutes the grounds for integrative negotiations on allocating control and accountability in the newly emerging work systems.

In relation to AI, a pertinent example is Kim, Glaeser, Hillis, Kominers, and Luca's (2024) discussion of why restaurant inspectors preferred their own heuristiscs to AI-based recommendations for inspection targets, even though the latter were more accurate. This case illustrates the relevance of jointly considering rationalist and constructivist viewpoints for developing more effective work practices. Likewise, Lebovitz, Lifshitz-Assaf, and Levina (2022) reported that physicians would sincerely engage with AI-based diagnostic tools only after challenging the logic of the systems based on their own expertise. Lebovitz et al. (2021) also described how decision-makers in a hospital struggled to verify the accuracy of AI tools used for diagnostic tasks. They tried to understand how “ground truths” were established in

the training and validation of the tools and how this process compared to their own practice of ensuring the most accurate decisions. Key to physicians' frustration was that their concept of expert judgment, based on sensemaking and intuition, conflicted with the technology developers' purely rationalist understanding. These latter examples show how the lack of perspective taking between AI developers and users impedes effective negotiation of control and accountability and as a consequence effective use of a new AI systems. This discussion leads to the fifth proposition:

Proposition 5. To the extent that stakeholders engage in extensive information sharing aided by perspective taking, an integrative negotiation strategy is more likely.

Turning to managing power differentials as a second lever for promoting an integrative negotiation strategy, one must first acknowledge abundant evidence that more powerful stakeholders often manage to obtain more satisfactory results for themselves at the cost of other less powerful stakeholders (Brett & Thompson, 2016; Thompson et al., 2010). Regarding AI development and use, this imbalance can be illustrated by the quarrel between Tesla and the National Highway Traffic Safety Administration (NHTSA) over Tesla's Autopilot. In the wake of several severe accidents, NHTSA's attempts to gain more knowledge on the exact functioning of this system were stymied; indeed, NHTSA eventually bent to Elon Musk's pressures and removed a prominent human factors expert from its board (Ross, 2023). Uber is another negative example as the company resists to acknowledge its role as an employer and refuses to disclose the algorithms used in driver-customer matching and performance evaluation (Möhlmann, Zalmanson, Henfridsson, & Gregory, 2021).

If stakeholders adopt more decentralized forms of governance, as postulated in our third proposition, this should help to increase powerful actors' willingness to accept accountability and to give up some of their control to others to help them fulfil their accountabilities (Anthony et al., 2023; Chhillar & Aguilera, 2022). Recent conceptual work on stakeholder governance has been driven by optimism that power struggles will subside once stakeholders

realize that egalitarian collaboration provides the best outcomes for all when tasks are highly interdependent (Bridoux & Stoelhorst, 2022). Decentralized stakeholder governance should thus facilitate frame alignment towards accepting accountability for mitigating AI risks as a shared norm (Grimm & Reinecke, 2024). Thereby, even big players such as the organizations that possess large data sets and the computational capability for training complex models might be swayed to accept accountability for the quality of data and algorithms they provide and overcome concerns about losing competitive advantage by making systems more transparent to users and regulators (de Laat, 2018; Faraj et al., 2018; Jacobides et al., 2021).

Moreover, worries regarding control-accountability alignment by less powerful actors such as individual AI users and developers are likely to be heard more if accountability is a shared concern. Not only would AI users such as the physicians in the studies of Lebovitz and colleagues (2021, 2022) have a stronger voice in stakeholder negotiations, but also individual AI developers. Suchman (2002: 94) has argued that in order to settle what she calls located accountabilities, developers "must give up control over technology design (which is in any case illusory) and see themselves as entering into an extended set of working relations for which the question at each next turn becomes: How do we proceed in a responsible way?" If accountability is a shared norm by all, for instance the informal and highly personalized strategies used by AI developers in a study by Hagtvedt, Harvey, Demir-Caliskan & Hagtvedt (2024) to deliberately limit the functionality of new AI systems to keep human users in the loop might become accepted or even desired practice in order to mitigate AI risks (see also the discussion leading up to proposition 2 and Table 1). Establishing such practices also mirrors discussions on accuracy-explainability tradeoffs for explainable AI (Kim & Doshi-Velez, 2021). The assumption that more complex and thus more opaque systems are always more accurate has been challenged by pointing to cases where simpler models were as accurate, and by highlighting the business interests involved in selling complex models to customers who cannot verify accuracy claims (Rudin, 2019).

These examples illustrate how more integrative negotiations of control-accountability alignment can be fostered despite strong power differentials among stakeholders if accountability for mitigating AI risks is established as a shared norm. Chhillar and Aguilera (2022) have emphasized that norms guiding self-regulation of stakeholder decision-making may be more effective than hard law, because law enforcement is very difficult to achieve in highly entangled processes, such as those characteristic of AI development and use. However, regulatory action can help to build an 'accountability culture' in which all stakeholders are committed to mitigating AI risks (NIST, 2023), for instance by developing indicators for such a culture within a regulatory framework of self-regulation, as has happened in a number of high-risk industries (Grote, 2012; Kirwan, Hale & Hopkins, 2002; Majumdar & Markus, 2001; May, 2007). This discussion leads to our final proposition:

Proposition 6. To the extent that accountability for mitigating AI risks is a shared norm among stakeholders, an integrative negotiation strategy is more likely even in the case of large power differentials between stakeholders.

DISCUSSION

Our theory aims to answer the fundamental question of how the principle of control-accountability alignment can be upheld for AI development and use to mitigate AI risks. Core to our theorizing about the processes involved in developing and using AI has been that, compared to other technologies, there are new risks caused by the increasing autonomy and adaptive capabilities of AI. As AI systems learn from huge data sets and continuously change during their use, they become opaque even for their developers, and new interdependencies are created between developers and users. Faulty and biased decision-making in a wide variety of AI applications, uncontrolled use of private data, and unvetted use of information produced by generative AI bear witness to these new risks. Alignment between control and accountability as a fundamental way to contain risk in organizational processes is rendered more difficult because AI opacity and dynamism reduce human control and the entanglement

between development and use hinders unequivocal assignment of control and accountability. We have detailed how decisions on allocating control and accountability may acknowledge these new challenges and have postulated decentralized forms of stakeholder governance and integrative negotiation strategies as conducive to achieving control-accountability alignment.

We argue that by fostering egalitarian interactions among stakeholders based on a shared sense of accountability for keeping risks at bay in the highly complex and interdependent processes during AI development and use, a fuller understanding of the challenges involved in aligning control and accountability is developed, which can guide integrative negotiations aimed at sharing the burdens and benefits of AI. Lu, D'Agostino, Rudman, Ouyang, and Ho (2022) provide a good example for the postulated processes, where researchers worked with a public health agency to develop an AI-based Covid-19 contact tracing tool. Through extensive consultation, common ground was built, which led to using a model comprehensible to all stakeholders, rather than the originally envisioned slightly more accurate but also far more complex models.

Contributions

Foremost, our theory aims to contribute to mitigating AI risks by proposing how control and accountability can be aligned for AI users and developers and their senior management through stakeholder negotiations in decentralized forms of governance. We provide a framework that allows management researchers to not only describe and explain the impact of AI but to also examine organizational processes involved in shaping its development and use.

At the micro level, our theory speaks to the working conditions of AI users and developers. It has long been acknowledged, but not resolved, that users of technology frequently find themselves in situations where they must incorporate a technical system into their work processes without sufficiently understanding the system nor having appropriate means to influence it in line with performance goals (Parker & Grote, 2022). Such situations are demotivating and stressful and can inhibit learning. However, research on the working

conditions of technology developers is scarce (Anthony et al., 2023). Our theory emphasizes that AI risk mitigation requires a closer look at whether AI developers themselves have sufficient means to control AI development and how they may be able to compensate for partially losing control as AI systems become more autonomous and dynamic. By recognizing the role of AI developers more explicitly, simply shifting blame for inadequate performance from users to developers is avoided. Only when adequate consideration is given to whether developers can adequately bear accountability, they will be motivated to strive for effective and safe AI systems. Moreover, by examining changing working requirements for AI developers, it may become apparent that they require additional skills, especially concerning socio-technical design thinking and interaction with multiple stakeholders necessary to establish productive feedback processes.

At the meso level, our portrayal of the intricate relationships between control and accountability in AI development and use can contribute to a better understanding of how control and accountability can be aligned across different actors to manage the "many hands" problem in highly interdependent tasks. We were intrigued to find that control and accountability have largely been treated in separate research streams in the organizational literature (Merchant & Otley, 2007; Sitkin at al, 2020). The highly entangled interactions between multiple actors across several organizations involved in AI development and use can be an interesting testbed for investigating how controlling others through holding them accountable can be more or less effective depending on how much control over actual work processes and outcomes those others have. Furthermore, the case of AI development and use speaks to fundamental issues of sharing control and accountability amongst various actors. For instance, AI developers may be held accountable for actions of AI users when AI systems cannot be rendered sufficiently transparent and predictable for AI users; and senior managers may be held accountable when AI systems are impenetrable to developers and yet developers are not encouraged and enabled to develop complementary mechanisms for risk mitigation or

even voice concerns (Lukpat, 2024). We outline organizational mechanisms that can support such distributed forms of control-accountability alignment and thereby inform research on processes involved in truly sharing rather than diffusing responsibility.

At the macro level, our propositions for integrative stakeholder negotiations in decentralized forms of governance may enrich strategic considerations for the management of technologies in organizations and broader concerns of effective stakeholder governance. By delineating how negotiations on control and accountability require extensive information sharing and perspective taking amongst multiple and differently situated stakeholders, our theory encourages a deeper appreciation of the interplay between rationalist and constructivist perspectives involved in managing the uncertainties of emerging technologies (Kapoor & Klueter, 2021). By addressing the fundamental problem of how powerful stakeholders may be motivated to relinquish some of their power to achieve better results for all, we also add to the growing literature on AI governance (Chhillar & Aguilera, 2022; Falco et al., 2021; Gasser & Almeida, 2017; Wirtz et al., 2022). In line with recent literature on stakeholder governance, we argue that shared governance among all stakeholders can be effective, especially if accountability can be agreed upon as a common norm and powerful actors realize that thereby reputation and other risks can be reduced (Bridoux & Stoelhorst, 2022).

Future research

By putting our propositions to the test, future research will be instrumental for shaping AI development and use in ways that capitalize on AI's potential for improved performance in a wide array of businesses, while keeping its risks at bay. Research may cut across levels of analysis to examine the impact of different options for control-accountability alignment for different stakeholders or the effectiveness of different mechanisms for integrative negotiation strategies we have proposed. Furthermore, a range of questions can be addressed that follow from the micro, meso, and macro considerations we have outlined above.

At the micro level, research could delve into the daily work of AI developers to better understand how they deal with the opportunities and challenges involved in creating AI systems. Such research could follow the example of ethnographic studies mentioned earlier (Hadtvedt et al., 2024; Jain, 2023) to learn more about developers' strategies to stay in control themselves and to also help AI users to stay in control. As an example, a study by Myers (2023) unravelled some of the power dynamics that AI developers face and how they cope with them. She found that developers managed to address worker concerns in their designs, even against the interests of those workers' managers. For instance, they set up a system for workers to report operational problems in a way that allowed them to also communicate issues they themselves wanted solved. Adding to this emerging body of knowledge can help to inform work design for AI developers, drawing on the immense research on what job characteristics are key for designing motivating work and improving workers' performance and well-being (Parker, 2014).

Besides establishing a new focus on the working conditions of AI developers, research should continue to address the work by AI users as it is transformed through new opportunities for augmenting, but also automating, human capabilities. Most fundamentally, such research concerns the allocation of functions between human users and technical systems which with their increasing agency become stakeholders in their own right (Choudhary et al., 2023; Johnson & Verdicchio, 2019). Baird and Maruping (2021) have developed an elaborate conceptual model of how negotiations are part of dynamically delegating tasks between human actors and AI systems where capabilities, roles, and preferences along with coordination requirements and liabilities are assessed by both, putting the AI at the same level as the human actor. Empirically testing their assumptions with the issue of aligning control and accountability for both in mind will be a fascinating first step towards discussing possibilities for accountability residing in technical systems.

At the meso level, research may address the dynamics involved in aligning control and accountability for individual actors as well as across actors. For instance, Frink et al.'s (2008) multilevel model of accountability could be employed to trace how developments in AI regulatory systems as they are currently underway in many countries, will affect both organizational accountability and, through the control mechanisms companies choose to answer to their new accountabilities, also individuals' accountability and control. Such research could also take different kinds of organizational arrangements into account. In our theorizing we have assumed a simple situation of one company developing technology for other companies, which provides a basic structure for assigning control and accountability. However, AI development and use might happen in many other settings, including nascent firms, freelancers, organizations that possess data and computing power, consumers as users rather than firms, which all imply different sets of actors and different fora these actors are accountable to. Moreover, we have not considered challenges and opportunities posed by different environments for how control and accountability might be aligned, for instance market pressures or legal requirements. Effective governance is highly dependent on such factors and our propositions could therefore be expanded to include environmental contingencies (Aguilera et al., 2008).

Testing our propositions on the positive effects of allocating accountability to AI developers for the outcomes of using AI systems can provide valuable insights into new forms of distributed control and accountability not addressed to date. Cutting across levels of analysis, such research could examine different kinds of accountability, which we have not done to keep the complexity of our propositions manageable. For instance, in their research on human-centred system design for Internet of Things technologies, Boos et al. (2013) distinguished three kinds of accountability: *Visibility* refers to the demand to provide an intelligible account of one's own actions to other actors (Suchman, 2007); *responsibility* concerns an actor's duties and obligation to perform certain activities, which is determined by

the actor's formal work role and the specific allocations of functions between different human actors and between humans and technology (Lee, 2006); and *liability* relates to an actor's requirement to answer to law, regulations, and contracts (Nissenbaum, 1996).

These three kinds of accountability may not necessarily be assigned to the same actor. Often, visibility and responsibility will go together because being responsible for fulfilling a required task requires actors to show others what actions they have taken to do so. An intricate issue regarding the linkage between visibility and responsibility has been pointed out by Leonardi and Treem (2020) which they term the transparency paradox: Actors may provide a flurry of information about their behavior that makes identifying relevant information for judging the appropriateness of their behavior more difficult. Liability may often be assigned to third actors, for instance to managers who delegated tasks to members of their team, or even to the firm as a whole. Furthermore, the three kinds of accountability interact with each other: Properly aligning control with both visibility and responsibility demands to ensure effective workflows will reduce the likelihood that liability claims materialize. Distinguishing different kinds of accountability can also advance discussions on whether humans remain ultimately accountable even for highly agentic AI systems. Visibility and responsibility may eventually be assigned to AI systems, while liability rests with humans.

At the macro level, our propositions on integrative stakeholder negotiations may provide a starting point for exploring decentralized forms of stakeholder governance further. Recently, several theoretical lenses have been proposed that all emphasize collaboration among stakeholders outside of classical hierarchies in organizations and the necessity for participation and dialogue to develop a shared purpose, to promote responsible innovation, and to prevent exploitation of common-pool resources (Bacq & Aguilera, 2022; Bridoux & Stoelhorst, 2022; McGahan, 2023). As Chhillar and Aguilera (2022) have shown, AI governance can benefit from knowledge about different accountability mechanisms, such as

financial incentives, legal requirements, and social stigma, and their effects on stakeholder compliance. Investigating such effects at the fine-grained level of stakeholder negotiations proposed in our theory may add empirical depth to emerging new forms of stakeholder governance and guide practitioners in bringing effective risk management of AI to life.

A crucial and overarching consideration for future research concerns the role of regulators in mitigating AI risks. A multitude of approaches for regulating AI exist, ranging from industrial standards to binding law and from norms to constrain AI development and use to norms that enable innovation and reduce power imbalances among stakeholders (Gasser, 2024). Regulators have very different functions depending on which regulatory regimes (May, 2007) are chosen, with different requirements and opportunities for how their own control and accountability can be aligned in the quest for safer AI development and use. Our focus has been on the interactions between stakeholders directly involved in AI development and use and how these interactions may be shaped to foster risk mitigation. However, these interactions are strongly influenced by the chosen regulatory regime. Whether indeed regimes can be successful at mitigating AI risk that emphasize self-regulation among stakeholders, as we have in our theorizing, and permit learning and exploration by "flexible rules" (Grote, 2024) in support of "tentative governance" (Gasser, 2024), is a question at the core of current efforts to tame AI in the service of society.

CONCLUSION

The risks posed by AI seem to be on everyone's mind, from casual AI users to nations. By focusing on two fundamental concepts in management research—control and accountability—we propose that the risks of highly complex, opaque, and dynamic advanced AI systems can be mitigated through integrative stakeholder negotiations aimed at equitable and effective control-accountability alignment. New research opportunities abound to empirically validate and expand our reasoning for how AI's fast-growing agency can be tamed to make AI safer and more useful for all.

REFERENCES

Aguilera, R. V., Filatotchev, I., Gospel, H., & Jackson, G. 2008. An organizational approach to comparative corporate governance: Costs, contingencies, and complementarities. ***Organization Science***, 19: 475–492.

Amis, J., Barney, J., Mahoney, J. T., & Wang, H. 2020. Why we need a theory of stakeholder governance—and why this is a hard problem. ***Academy of Management Review***, 45: 499–503.

Anthony, C., Bechky, B. A., & Fayard, A. (2023). "Collaborating" with AI: Taking a system view to explore the future of work. ***Organization Science***, 34: 1672-1694.

Asatiani, A., Malo, P., Nagbol, P. R., Penttinen, E., Rinta-Kahila, T., & Salovaara, A. 2021. Sociotechnical envelopment of artificial intelligence: An approach to organizational deployment of inscrutable artificial intelligence systems. ***Journal of the Association for Information Systems****,* 22: 325–352.

Babic, B., Gerke, S., Eveniou, T., & Cohen, I. G. 2019. Algorithms on regulatory lockdown in medicine. ***Science****,* 366(6470): 1202–1204.

Bacq, S., & Aguilera, R. V. 2022. Stakeholder governance for responsible innovation: A theory of value creation, appropriation, and distribution. ***Journal of Management Studies***, 59: 29–60.

Bailey, D. E., & Barley, S. R. 2020. Beyond design and use: How scholars should study intelligent technologies. ***Information and Organization***, 30: 100286.

Bainbridge, L. 1983. Ironies of automation. Automatica*,* 19: 775–779.

Baird, A., & Maruping, L. M. 2021. The next generation of research on IS use: A theoretical framework of delegation to and from agentic IS artifacts. ***MIS Quarterly****,* 45: 315–341.

Balasubramanian, N., Ye, Y., & Xu, M. 2022. Substituting human decision-making with Machine Learning: Implications for organizational learning. ***Academy of Management Review*****,** 47: 448–465.

Bartsch, S., Milani, V., Adam, M., & Benlian, A. (2024). Algorithmic accountability: What does it mean for AI developers and how does it affect AI development projects. ***Proceedings of the 57th Hawaii International Conference on System Sciences.***

Bengio, Y. 2024. Managing extreme AI risks amid rapid progress. ***Science***, 384: 843–845.

Berente, N., Gu, B., Recker, J., & Santhanam, R. 2021. Managing Artificial Intelligence. ***MIS Quarterly,*** 45: 1433–1450.

Billings, C. E. 1997. ***Aviation automation: The search for a human-centered approach***. Mahwah, NJ: Lawrence Erlbaum.

Boos, D., Grote, G. & Guenter, H. 2013. Controllable accountabilities. The Internet of Things and its challenges for organisations. ***Behaviour & Information Technology,*** 32: 449–467.

Bovens, M. 2007. Analysing and assessing accountability: A conceptual framework. ***European Law Journal,*** 13: 447–468.

Brehmer, B. 1992. Dynamic decision making—human control of complex systems. ***Acta Psychologica,*** 81: 211–241.

Brett, J., & Thompson, L. 2016. Negotiation. ***Organizational Behavior and Human Decision Processes***, 136: 68–79.

Bridoux, F., & Stoelhorst, J. W. 2022. Stakeholder governance: Solving the collective action problems in joint value creation. ***Academy of Management Review***, 47: 214-236.

Brock, J., & von Wangenheim, F. 2019. Demystifying AI: What digital transformation leaders can teach you about realistic artificial intelligence. ***California Management Review,*** 61: 110–134.

Burton, S., Habli, I., Lawton, T., McDermid, J., Morgan, P., & Porter, Z. 2020. Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective. ***Artificial Intelligence***, 279: 103201.

Carver, C. S., Scheier, M. F. 1990. Origins and functions of positive and negative affect: A control-process view. ***Psychological Review***, 97: 19–35.

Castelvecchi, D. 2016. The blackbox of AI. ***Nature,*** 538(7623): 21–23.

Challenger, R., Clegg, C. W., & Shepherd, C. 2013. Function allocation in complex systems: reframing an old problem. ***Ergonomics,*** 56: 1051–1069.

Chan, A., Bradley, H., & Rajkumar, N. 2023. Reclaiming the digital commons: A public data trust for training data. arXiv:2303.09001v2

Chen, Y., Richter, J. I., & Patel, P. C. (2021). Decentralized governance of digital platforms. ***Journal of Management***, 47: 1305–1337.

Chhillar, D., & Aguilera, R. V. 2022. An eye for artificial intelligence: Insights into the governance of artificial intelligence and vision for future research. ***Business & Society***, 61: 1197–1241.

Choudhary, V., Marchetti, A., Shresta, Y. R., & Purunam, P. 2023. Human-AI ensembles: When can they work? ***Journal of Management.***

Churchman, D. (2019). Negotiation. In: ***The Palgrave Encyclopedia of Peace and Conflict Studies***. Cham, Switzerland: Palgrave Macmillan.

Clegg, C. W. 2000. Sociotechnical principles for system design. ***Applied Ergonomics***, 31: 463–477.

Cobbe, J., Veale, M., & Singh, J. 2023. Understanding accountability in algorithmic supply chains. In ***Proceedings of FAccT '23, June 12–15, 2023, Chicago, IL, USA***: 1186–1197.

Crowston, K., & Bolici, F. 2020. Impacts of the use of machine learning on work design. ***8th International Conference on Human-Agent Interaction***, November 10–13, 2020, Virtual Event, NSW, Australia. https://doi.org/10.1145/3406499.3415070

Curhan, J. R., Overbeck, J. R., Cho, Y., Zhang, T., & Yang, Y. 2022. Silence is golden: Extended silence, deliberative mindset, and value creation in negotiation. ***Journal of Applied Psychology***, 107: 78–94.

De Dreu, C. K. W., Koole, S. L., & Steinel, W. 2000. Unfixing the fixed pie: A motivated information-processing approach to integrative negotiation. ***Journal of Personality and Social Psychology***, 79: 975–987.

De Dreu, C. K. W., Beersma, B., Stroebe, K., & Euwema, M. C. 2006. Motivated information processing, strategic choice, and the quality of negotiated agreement. ***Journal of Personality and Social Psychology***, 90: 927–943.

de Laat, P. B. 2018. Algorithmic decision-making based on machine learning from big data: Can transparency restore accountability? ***Philosophy of Technology,*** 31: 525–541.

Diakopoulos, N. 2016. Accountability in algorithmic decision making. ***Communications of the ACM,*** 59: 56–62.

Dolata, M., & Crowston, K. 2023. Making sense of AI systems development. ***IEEE Transactions on Software Engineering.***

Eisenhardt, K. M. 1989. Agency theory: An assessment and review. ***Academy of Management Review***, 14: 57–74.

Elish, M. C. 2019. Moral crumple zones: Cautionary tales in human-robot interaction. ***Engaging Science, Technology, and Society,*** 5: 40–60.

Endsley, M. R. 2017. From here to autonomy: Lessons learned from human-automation research. ***Human Factors***, 59: 5–27.

Endsley, M. R. 2023. Supporting human-AI teams: Transparency, explainability, and situation awareness. ***Computers in Human Behavior***, 140: 107574.

European Union 2023. Artificial Intelligence Act. Retrieved from https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf

Falco, G., Shneiderman, B., Badger, J., Carrier, R., Dahbura, A., Danks, D. et al. 2021. Governing AI safety through independent audits. ***Nature Machine Intelligence,*** 3: 566–571.

Faraj, S., Pachidi, S., & Sayegh, K. 2018. Working and organizing in the age of the learning algorithm. ***Information and Organization,*** 28: 62–70.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. 2018. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. ***Minds and Machines,*** 28: 689–707.

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. 2021. How to design AI for social good: Seven essential factors. In L. Floridi (ed.), ***Ethics, governance, and policies in Artifical Intelligence*** (pp. 125–151). Cham, Switzerland: Springer Nature.

Frink, D. D., Hall, A. T., Perryman, A. A., Ranft, A. L., Hochwarter, W. A., Ferris, G. R., & Royle, M. T. 2008. Meso-level theory of accountability in organizations. ***Research in Personnel and Human Resources Management***, 27: 177–245.

Future of Life Institute 2023. Pause giant AI experiments: An open letter. Retrieved from https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Galinsky, A. D., Maddux, W. W., Gilin, D., & White, J. B. 2008. Why it pays to get inside the head of your opponent. The differential effects of perspective-taking and empathy in negotiations. ***Psychological Science***, 19: 378–384.

Gasser, U. 2024. Governing AI with intelligence. ***Issues in Science and Technology*** (May 21, 2024). https://doi.org/10.58875/AWJG1236

Gasser, U., & Almeida, V. A. F. 2017. A layered model for AI governance. ***IEEE Internet Computing,*** 21: 58–62.

Glassman, E. L., Gu, Z., & Kummerfeld, J. K. 2024. AI-resilient interfaces. arXiv:2405.08447v1

Green, S.G. & Welsh, M.A. 1988. Cybernetics and dependence–reframing the control concept. ***Academy Management Review,*** 13: 287–301.

Grimm, J., & Reinecke, J. (2024). Collaborating on the edge of failure: Frame alignment across multiple interaction arenas in multi-stakeholder partnerships. ***Academy of Management Journal***, 67: 956–990.

Grote, G. 2012. Safety management in different high-risk domains–All the same? ***Safety Science,*** 50: 1983–1992.

Grote, G. 2024. Uncertainty regulation in high-risk organizations: Harnessing the benefits of flexible rules. In J. Le Coze & B. Journé (Eds.), ***Compliance and initiative in the production of safety***: 13–20). Cham, Switzerland: Springer.

Grote, G., Ryser, C., Wäfler, T., Windischer, A., & Weik, S. 2000. KOMPASS: a method for complementary function allocation in automated work systems. ***International Journal of Human-Computer Studies,*** 52: 267-287.

Habli, I., Lawton, T., & Porter, Z. 2020. Artificial intelligence in health care: Accountability and safety. ***Bulletin of the World Health Organization***, 98: 251-256.

Hagtvedt, L. P., Harvey, S., Demir-Caliskan, O., & Hagtvedt, H. (2024). Bright and dark imagining: How creators navigate moral consequences of developing ideas for artificial intelligence. ***Academy of Management Journal***, published online first.

Hall, A. T., Frink, D. D., Buckley, M. R. 2017. An accountatbility account: A review and synthesis of the theoretical and empirical research on felt accountability. ***Journal of Organizational Behavior***, 38: 204–224.

Hall, A. T., Frink, D. D., Ferris, G. R., Hochwarter, W. A., Kacmar, C. J., & Bowen, M. G. 2003. Accountability in human resource management. In C. A. Schriesheim & L. L. Neider (Eds.), ***New directions in Human Resource Management***: 29–64. Greenwich, Conn.: IAP.

Hall, A. T., Royle, M. T., Brymer, R. A., Perrewé, P. L., Ferris, G. R., & Hochwarter, W. A. 2006. Relationship between felt accountability as a stressor and strain reactions: The neutralizing role of autonomy across two studies. ***Journal of Occupational Health Psychology***, 11: 87–99.

Hardy, C., Maguire, S., Power, M., & Tsoukas, H. 2020. Organizing risk: Organization and management theory for the risk society. ***Academy of Management Annals,*** 14: 1032–1066.

Hartmann, M., & Beane, M. (2024). Haunted adoption: How applied experts can drive organizations to incorporate technology without strong managerial support. Available at SSRN 4925724.

Hern, A. (2024). TechScape: What we learned from the global AI summit in South Korea. The Guardian, May 28, 2024. Retrieved from https://www.theguardian.com/technology/article/2024/may/28/techscape-ai-global-summit

Holford, W. D. 2022. An ethical inquiry of the effect of cockpit automation on the responsibilities of airline pilots: Dissonance or meaningful control? ***Journal of Business Ethics,*** 176: 141–157.

Hollnagel, E., & Woods, D. D. 2005. ***Joint cognitive systems: Foundation of cognitive systems engineering***. Taylor and Francis, New York.

Jarvenpaa, S. L., & Välikangas, L. 2020. Advanced technology and endtime in organizations: A doomsday for collaborative creativity? ***Academy of Management Perspectives,*** 34: 566–584.

Jacobides, M. G., Brusoni, S., & Candelon, F. 2021. The evolutionary dynamics of the Artificial Intelligence ecosystem. ***Strategy Science,*** 6: 412–435.

Jain, P. 2023. The yellow brick road to artificial intelligence: An empirical study of developers developing artificial intelligent conversational socialbots. Doctoral dissertation, Stanford University, Department of Management Science and Engineering.

Jamieson, G. A., Skraaning, G. & Joe, J. 2022. The B737 MAX 8 accidents as operational experiences with automation transparency. ***IEEE Transactions on Human-Machine Systems,*** 52: 794–797.

Jasperson, J., Carte, T. A., Saunders, C. S., Butler, B. S., Croes, H. J. P., & Zheng, W. 2002. Power and Information Technology research: A metatriangulation review. ***MIS Quarterly,*** 26: 397–459.

Johnson, D. G., & Verdicchio, M. 2019. AI, agency and responsibility: The VW fraud case and beyond. ***AI & Society***, 34: 639–647.

Kane, G. C., Young, A. G., Majchrzak, A., & Ransbotham, S. 2021. Avoiding an oppressive future of machine learning: A design theory for emancipatory assistants. ***MIS Quarterly,*** 45: 371–396.

Kapoor, R., & Klueter, T. 2021. Unbundling and managing uncertainty surrounding emerging technologies. ***Strategy Science,*** 6: 62–74.

Kellogg, K. C., Valentine, M., & Christin, A. 2020. Algorithms at work: The new contested terrain of control. ***Academy of Management Annals,*** 14: 366–410.

Kemp, A. 2023. Competitive advantage through artificial intelligence: Toward a theory of situated AI. ***Academy of Management Review***.

Kim, B. & Doshi-Velez, F. 2021. Machine learning techniques for accountability. ***AI Magazine***, Spring 2021, 47–52.

Kim, H., Glaeser, E. L., Hillis, A., Kominers, S. D., & Luca, M. 2024. Design authority and the returns to algorithms. ***Strategic Management Journal***, 45: 619–648.

Kirwan, B., Hale, A.R., Hopkins, A. (Eds.). 2002. ***Changing regulation: Controlling hazards in society***. Oxford: Pergamon.

de Laat, P. B. 2018. Algorithmic decision-making based on machine learning from big data: Can transparency restore accountability? ***Philosophy of Technology,*** 31: 525–541.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. 2021. What do we want from Explainable Artificial Intelligence (XAI)? - A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. ***Artificial Intelligence***, 296: 103473.

Lebovitz, S., Levina, N., & Lifshitz-Assaf, H. 2021. Is AI ground truth really "true"? The dangers of training and evaluating AI tools based on experts' know-what. ***MIS Quarterly***, 45: 1501–1525.

Lebovitz, S., Lifshitz-Assaf, H., & Levina, N. 2022. To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. ***Organization Science,*** 33: 126–148.

Lee, J.D. 2006. Human factors and ergonomics in automation design. In G. Salvendy, ed. ***Handbook of Human Factors and Ergonomics*** (3rd ed.) (pp. 1570–1596). New Jersey: Wiley & Sons.

Leonardi, P. M. 2012. Materiality, sociomateriality, and socio-technical systems: What do these terms mean? How are they different? Do we need them. In P. M. Leonardi, B. A. Nardi, & J. Kallinikos (Eds.), ***Materiality and organizing: Social interaction in a technological world***. Oxford, UK: Oxford University Press.

Leonardi, P. M., & Barley, S. R. 2010. What's under construction here? Social action, materiality, and power in constructivist studies of technology and organizing. ***Academy of Management Annals,*** 1: 1–51.

Leonardi, P. M., & Treem, J. W. (2020). Behavioral visibility: A new paradigm for organizational studies in the age of digitization, digitalization, and datafication. ***Organizational Studies***, 41: 1601–1625.

Lerner, J. S., Tetlock, P. E. 1999. Accounting for the effects of accountability. ***Psychological Bulletin***, 125: 255–275.

Lindebaum, D., Vesa, M., & Den Hond, F. 2020. Insights from the machine stops to better understand rational assumptions in algorithmic decision-making and its implications for organizations. ***Academy of Management Review***, 45: 247–263.

Lu, L., D'Agostino, A., Rudman, S. L., Ouyang, D., & Ho, D. E. 2022. Designing accountable health care algorithms: Lessons from Covid-19 contact tracing. ***NEJM Catalyst***, 3.

Lukpat, A. (2024). AI employees fear they aren´t free to voice their concerns. Wall Street Journal, June 4, 2024. Retrieved from https://www.wsj.com/tech/ai/ai-employees-fear-they-arent-free-to-voice-their-concerns-abd38606?st=qhdolkzyxnmet6g&reflink=desktopwebshare_permalink

Macrae, C. 2022. Learning from the failures of autonomous and intelligent systems: Accidents, safety, and sociotechnical sources of risk. ***Risk Analysis***, 42: 1999-2025.

Magee, J. C., & Galinsky, A. D. 2008. Social hierarchy: The self-reinforcing nature of power and status. ***Academy of Management Annals***, 2: 351–398.

Majumdar, S. K., & Marcus, A. A. 2001. Rules versus discretion: The productivity consequences of flexible regulation. ***Academy of Management Journal,*** 44: 170–179.

Makarius, E. E., Mukherjee, D., Fox, J. D., & Fox, A. K. 2020. Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. ***Journal of Business Research,*** 120: 262–273.

Markus, M. L. 2017. Datification, organizational strategy, and IS research: What's the score? ***Journal of Strategic Information Systems***, 26: 233–241.

May, P.J., 2007. Regulatory regimes and accountability. ***Regulation & Governance***, 1: 8–26.

McGahan, A. M. 2023. The new stakeholder theory on organizational purpose. ***Strategy Science***, 8: 245–255.

Merchant, K. A., & Otley, D. T. 2007. A review of the literature on control and accountability. In C. S. Chapman, A. G. Hopwood & M. D. Shields (Eds.), ***Handbook of management accounting research***: 785–802. Amsterdam: Elsevier.

Milmo, D. (2024). OpenAI putting 'shiny products' above safety, says departing researcher. The Guardian, May 18, 2024- Retrieved from 'https://www.theguardian.com/technology/article/2024/may/18/openai-putting-shiny-products-above-safety-says-departing-researcher

Mittelstadt, B., Russell, C., & Wachter, S. 2019. Explaining explanations in AI. Proceedings of FAT 19, Conference on Fairness, Accountability, and Transparency, January 2019, Atlanta, USA.

Möhlmann, M., Zalmanson, L., Henfridsson, O., & Gregory, R. W. 2021. Algorithmic management of work on online labor platforms: When matching meets control. ***MIS Quarterly,*** 45: 1999–2022.

Murray, A., Rhymer, J., & Sirmon, D. G. 2021. Human and technology: Forms of conjoined agency in organizations. ***Academy of Management Review,*** 46: 552–571.

Myers, J. 2023.When Big Brother is benevolent: How technology developers navigate power dynamics among users to elevate worker interests. ***Academy of Management Discoveries.***

Norman, D., & Euchner, J. 2023. Design for a better world. ***Research-Technology Management***, 66: 11-18.

Nissenbaum, H. 1996. Accountability in a computerized society. ***Science and Engineering Ethics,*** 2: 25–42.

NIST 2023. Artificial Intelligence risk management framework (AI RMF 1.0). Retrieved from https://doi.org/10.6028/NIST.AI.100-1

Oliver, N., Calvard, T., & Potocnik, K. 2017. Cognition, technology and organizational limits: Lessons from the Air France 447 disaster. ***Organization Science***, 28: 597–780.

Orlikowski, W. J., & Scott, S. V. 2008. Sociomateriality: Challenging the separation of technology, work and organization. ***Academy of Management Annals,*** 2: 433–474.

Ouchi, W. G. 1979. A conceptual framework for the design of organizational control mechanisms. ***Management Science***, 25: 833–48.

Parker, S. K. 2014. Beyond motivation: Job and work design for development, health, ambidexterity, and more. ***Annual Review of Psychology,*** 65: 661–691.

Parker, S. K., Atkins, P. W., & Axtell, C. M. (2008). Building better workplaces through individual perspective taking: A fresh look at a fundamental human process. ***International Review of Industrial and Organizational Psychology***, 23: 149–196.

Parker, S. K., & Grote, G. 2022. Automation, algorithms, and beyond: Why work design matters more than ever in a digital world. ***Applied Psychology: An International Review***, 71: 1171–1204.

Raisch, S. & Fomina, K. 2023. Combining human and artificial intelligence: Hybrid problem-solving in organizations. ***Academy of Management Review***.

Raisch, S., & Krakowski, S. 2021. Artificial intelligence and management: The automation-augmentation paradox. ***Academy of Management Review,*** 46: 192–210.

Roose, K. 2023. A.I. belongs to the capitalists now. New York Times, Nov. 22, 2023. Retrieved from https://www.nytimes.com/2023/11/22/technology/openai-board-capitalists.html

Ross, P. E. 2023. A former pilot on why autonomous vehicles are so risky—Five questions for Missy Cummings. ***IEEE Spectrum***, June 2023.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. ***Nature Machine Intelligence,*** 1: 206–215.

Shneiderman, B. 2016. The dangers of fault, biased, or malicious algorithms requires independent oversight. ***Proceedings of the National Academy of Sciences,*** 113: 13538–13540.

Shresta, Y. R., Ben-Menahem, S. M., & von Krogh, G. 2019. Organizational decision-making structures in the age of artificial intelligence. ***California Management Review,*** 61: 66–83.

Sitkin, S. B., Long, C. P., & Cardinal, L. B. 2020. Assessing the control literature: Looking back and looking forward. ***Annual Review of Organizational Psychology and Organizational Behavior***, 7: 339–368.

Skinner, E. A. 1996. A guide to constructs of control. ***Journal of Personality and Social Psychology***, 71: 549–570.

Slota, S. C., Fleischmann, K . R. , Greenberg, S., Verma, N., Cummings, B., Li, L., & Shenefiel, C. 2023. Many hands make many fingers to point: Challenges in creating accountable AI. ***AI & Society***, 38: 1287–1299.

Srivastava, A. et al. 2023.Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv:2206.04615v3

Strich, F., Mayer, A., & Fiedler, M. 2021. What do I do in a world of artificial intelligence? Investigating the impact of substitutive decision-making AI systems on employees' professional role identity. ***Journal of the Association for Information Systems,*** 22: 304–324.

Suchman, L. A. 2002. Located accountabilities in technology production. ***Scandinavian Journal of Information Systems,*** 14: 91–105.

Suchman, L. 2007. ***Human-machine reconfigurations: Plans and situated actions.*** Cambridge: Cambridge University Press.

Taddeo, M., & Floridi, L. 2018. How AI can be a force for good. ***Science,*** 361: 751–752.

Teodorescu, M. H. M., Morse, L., Awwad, Y., & Kane, G. C. 2021. Failures of fairness in automation require a deeper understanding of Human-ML augmentation. ***MIS Quarterly,*** 45: 1483–1499.

Thompson, L. L., Wang, J., Gunia, B. C. 2010. Negotiation. ***Annual Review of Psychology***, 61: 491–515.

Van den Broek, E., Sergeeva, A., & Huysman, M. 2021. When the machine meets the expert: An ethnography of developing AI for hiring. ***MIS Quarterly***, 45: 1557–1580.

Van den Broek, E., Levina, N., & Sergeeva, A. 2022. In pursuit of data: Negotiating data tensions between data scientists and users of AI tools. ***Academy of Management Proceedings.***

Vincent, J. 2018. Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech. The Verge. Retrieved from https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai

Waardenburg, L., & Huysman, M. 2022. From coexistence to co-creation: Blurring boundaries in the age of AI. ***Information and Organization***, 32: 100432.

Waardenburg, L., Huysman, M., & Sergeeva, A. V. 2022. In the land of the blind, the one-eyed man is king: Knowledge brokerage in the age of learning algorithms. ***Organization Science***, 33: 59–82.

Weick, K. E., & Roberts, K. H. 1993. Collective mind in organizations: Heedful interrelating on flight decks. ***Administrative Science Quarterly***, 38: 357–381.

Weick, K. E., & Sutcliffe, K. M. 2001. ***Managing the unexpected***. San Francisco: Jossey-Bass.

The White House. 2023. Executive Order on the safe, secure, and trustworthy development and use of Artificial Intelligence. Washington, DC, October 30, 2023.

Wieringa, M. 2020. What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. ***Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency***, 1–18.

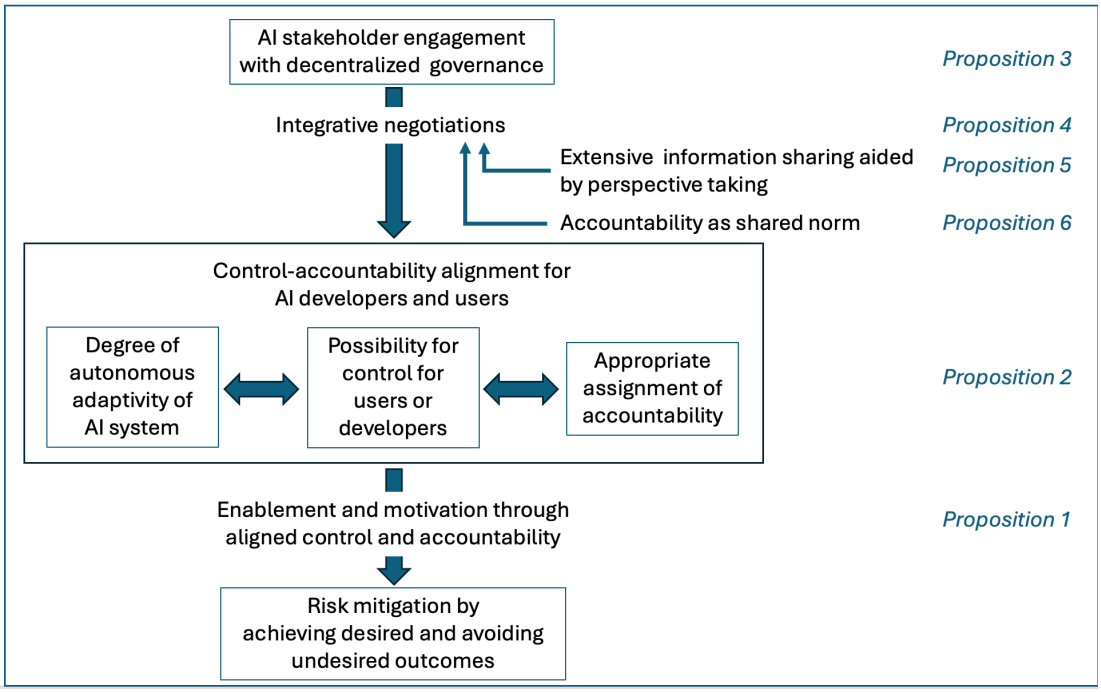
Wirtz, B. W., Weyerer, J. C., & Kehl, I. 2022. Governance of artificial intelligence: A risk and guideline-based integrative framework. ***Government Information Quarterly.***

Zuboff, S. 2019. ***The age of surveillance capitalism***. New York: PublicAffairs.

Table 1. Considerations for control-accountability alignment in AI development and use

AI system capabilities for autonomous adaptivity	**Low** (e.g., programmed system; trained system with few parameters)	**Medium** (e.g., trained system with many parameters, but frozen)	**High** (e.g., trained system with deep neural networks that continuously learns from new data)
Control of AI outcomes	**AI users** are in control if the system is explainable and leaves final decision-making to them.	**AI users** are at best partially in control if the system is explainable and still leaves certain decisions to them. **AI developers** are partially in control by defining and maintaining boundary conditions for system use.	**AI users** have no control and fully rely on the system as part of their work tasks, possibly aided by some explanations given by the system. **AI developers** are partially in control by intense testing and monitoring of system outcomes in line with an operating envelope.
Accountability for AI outcomes	**AI users** are accountable if conditions for their control have been established.	**AI developers** are accountable if conditions for their control have been established.	**AI developers** are accountable if conditions for their (partial) control have been established.
Control over AI system functioning	**AI developers** are in control if they have the decision power over ML techniques in line with the chosen system functionality.	**AI developers** are in control if they have the decision power over ML techniques in line with the chosen system functionality.	**AI developers** are partially in control by intense testing and monitoring of system functioning in line with an operating envelope.
Accountability for AI system functioning	**AI developers** are accountable if conditions for their control have been established.	**AI developers** are accountable if conditions for their control have been established.	**AI developers** are accountable if conditions for their control can be established; otherwise, **senior management of developers** is accountable.
Accountability for supporting organizational mechanisms	**Senior management of users and developers** are accountable for strengthening agency of AI users and developers.	**Senior management of users** are accountable for preventing control abuse by AI users. **Senior management of developers** are accountable for strengthening agency of AI developers.	**Senior management of users and developers** are accountable for endorsing organizational monitoring and feedback systems for continuous learning.

Figure 1. Conceptual model positioning our propositions



Author bios

Gudela Grote (ggrote@ethz.ch) is professor of work and organizational psychology at the Department of Management, Technology, and Economics, ETH Zürich, Switzerland. She received her PhD from the Georgia Institute of Technology. Her research examines the relation between technology, organization, and work, and the regulation of uncertainty by individuals and organizations.

Sharon K. Parker (s.parker@curtin.edu.au) is an Australian Research Council Laureate Fellow, Director of the Centre for Transformative Work Design, and a John Curtin Distinguished Professor of Organizational Behavior in the Faculty of Business and Law at Curtin University. She received her PhD from the University of Sheffield, UK. Her research focuses on job and work design, employee performance, proactive behaviour, organizational change, work stress, and quasi-experimental designs.

Kevin Crowston (crowston@g.syr.edu) is a Distinguished Professor of Information Science at the Syracuse University School of Information Studies. He received his Ph.D. from the Sloan School of Management, Massachusetts Institute of Technology. His research examines new ways of organizing made possible by information technology; theoretical characterizations of coordination problems and alternative methods for managing them; and design and empirical evaluation of systems to support people working together.