MIS Quarterly

THEORY & REVIEW

GUARDRAILS FOR HUMAN-AI ECOLOGIES: NORM-BASED COORDINATION AND DESIGN FOR PREDICTABILITY[1]

Thomas Grisold
WU Vienna University of Economics and Business, Vienna, AUSTRIA, and
University of St. Gallen, St. Gallen, SWITZERLAND {thomas.grisold@wu.ac.at}

Nicholas Berente
University of Notre Dame
South Bend, IN, U.S.A. {nberente@nd.edu}

Stefan Seidel
University of Cologne
Cologne, GERMANY {stefan.seidel@wiso.uni-koeln.de}

Human-AI ecologies involve human and AI-based agents that coordinate their interactions in part by following social norms. Social norms, therefore, are important for establishing the guardrails that ensure desirable interactions in a way that is consistent with essential values, such as human safety. Managing human-AI ecologies requires specifying norms to enable coordination in known situations but also allowing for the emergence of norms to enable coordination in unspecified, unstructured situations. We integrate predictive processing theory and social norm theory to explain how existing norms are enacted and reinforced based on agents' predictive models and how new norms emerge as agents update their predictive models in response to prediction errors in uncertain coordination scenarios. Rooted in this perspective, we develop a design theory that emphasizes design for predictability and propose a set of design principles for managers and developers to encode norms to evolve in human-AI ecologies, monitor outcomes, and intervene when necessary.

Keywords: Human-AI ecology, unstructured coordination, social norms, predictive processing, design principles

Introduction

Humans and AI-based agents—autonomous systems capable of learning—are increasingly enmeshed in shared environments and co-constitute what we refer to as *human-AI ecologies*. Human-AI ecologies range from hedonic applications like games to highly regulated ones like financial transactions and even safety-critical areas like transportation. In human-AI ecologies, humans and AI-based agents have complementary and evolving competences, which they enact and realize through joint interactions. Through these interactions, they mutually learn, creating ever-new situations (Berente et al., 2021; Lyytinen et al., 2021, Seidel et al., 2018). The potential for such ecologies is remarkable, as are the risks. It is essential for the managers of human-AI ecologies to understand the dynamic process of establishing and managing guardrails—or "zones of desirable behavior" (Gasser & Mayer-Schönberger, 2024)—for these ecologies as they evolve.

Just as any social collective, human-AI ecologies involve coordination among agents that repeatedly interact in a variety of situations (Claggett & Karahanna, 2018; Crowston, 1997; Faraj & Xiao, 2006). Coordinated interactions are enabled, in part, through norms (Hawkins et

[1] Ann Majchrzak was the accepting senior editor for this paper.

al., 2019; Ullmann-Margalit, 1977). Norms are social conventions that emerge to constitute shared expectations among social agents about joint interactions (Hawkins et al., 2019; Lewis, 1969; Ullmann-Margalit, 1977). We conceive of two types of norms: deep norms and surface norms. Deep norms represent abstract and stable social rules that apply in all kinds of situations and interactions (e.g., preserving safety, health, and rights). Surface norms are instantiations of deep norms that enable agents to learn and adjust their coordinated interactions (e.g., specific interactions that avoid violations of safety, health, and rights).

Existing research focuses on the emergence of norms among humans (Horne & Mollborn, 2020) or among AI-based agents in relatively well-defined task environments (Chopra et al., 2018). However, human-AI ecologies will increasingly require norm emergence in unstructured situations (Claggett & Karahanna, 2018; Faraj & Xiao, 2006) where these norms cannot be entirely specified in advance. Hence, a key issue for managers involves understanding and defining how norms emerge as humans *and* AI-based agents interact in human-AI ecologies involving unstructured and changing situations.

To address this issue, we integrate predictive processing theory with research on social norms to develop a design theory for norm-based coordination in human-AI ecologies. Predictive processing theory is an emerging paradigm in the computational neurosciences (Friston et al., 2017) that explains change and learning through uncertainty in both human agents *and* AI-based agents (Ciria et al., 2021; Clark, 2016). We develop a perspective that emphasizes design for predictability and generate a set of design principles, allowing for norm emergence to support coordination while averting undesired, potentially harmful norms and subsequent adverse situations (Kane et al., 2021). Specifically, we propose a set of design principles in two categories. First, we specify what is necessary so that *managers* can position human-AI ecologies within desired social boundaries. Human-AI ecologies should be designed in a way that allows managers to decide on and encode values into the ecology. To this end, managers, with the help of *developers,* must be enabled to implement values and other expectations via deep norms and selected surface norms to monitor and intervene when problematic surface norms emerge and to reinforce desirable surface norms as deep norms. Second, AI-based *agents* must be designed in ways that enable stable and predictable behavior when possible and flexible adjustments when necessary. To this end, agents should be enabled to foster predictability by communicating their own actions and sensing signals of other agents, to enact and adjust existing surface norms when possible, and to generate and reinforce new surface norms consistent with deep norms when necessary.

Coordination in Human-AI Ecologies

AI-based agents are increasingly enmeshed in complex social environments where they form collectives with dynamically evolving social relationships (Jennings et al., 2014; Lyytinen et al., 2021; Rahwan et al., 2019). We refer to these environments as *human-AI ecologies*. Human-AI ecologies are sociotechnical environments involving human and AI-based agents that interact with and learn from one another in flexible and emergent social relationships (Jennings et al., 2014; Lyytinen et al., 2021; Rahwan et al., 2019). Examples are shared traffic environments, virtual worlds and gaming environments, and automated logistics and warehousing.

We highlight three characteristics of human-AI ecologies: they take place in a shared environment, they are emergent, and they require coordination. First, as highlighted by the word *ecologies*, human-AI ecologies feature human and AI-based agents acting in a shared environment (Gibson, 1979; Lobo et al., 2018). Agent behaviors are enabled and constrained by a set of common environmental features. Agents exploit these features and attend to common cues in the environment (Browne & Garnham, 2022). For example, human and AI-operated vehicles navigate in shared traffic environments, using the same roads and attending to the same street signs (i.e., shared cues).

Second, human-AI ecologies are emergent: they are complex systems that change and evolve, which leads to unpredictable outcomes. While certain aspects can be predesigned, agents learn and change their behaviors and enter and exit ecologies over time (Jennings et al., 2014). This continuous variation creates uncertainty regarding interaction patterns. Hence, the development of human-AI ecologies and the agents within them cannot be fully determined or prescribed in advance (Lyytinen et al., 2021).

Third, human-AI ecologies require ongoing coordination. Coordination describes the process through which agents manage dependencies as they carry out activities (e.g., Claggett & Karahanna, 2018; Crowston, 1997). To coordinate, agents acquire information through communication and environmental cues (Faraj & Xiao, 2006; Okhuysen & Bechky, 2009). Coordination is often based on shared knowledge about respective skills, competences, and shared goals (Claggett & Karahanna, 2018).

Coordination can be more, or less, structured based on the information that is available to agents and the extent to which activities can be specified (Argote, 1982; Faraj & Xiao, 2006; Gittell, 2002). Therefore, coordination in human-AI ecologies can be structured or unstructured (e.g., Claggett & Karahanna, 2018). Structured coordination is prespecified and routinized, and agents know upfront what they should do in a given situation. Structured coordination is "programmed" through rules and procedures (Argote, 1982; Claggett & Karahanna, 2018). This implies that a developer can anticipate relevant situations and their associated coordination needs (Crowston, 1997) and define the relationships between agents in advance (Argote, 1982). Structured coordination is associated with low uncertainty and is linked to formal strategies of control (Claggett & Karahanna, 2018).

In contrast, unstructured coordination occurs when agents do not have predefined knowledge of how they should interact (Claggett & Karahanna, 2018; Faraj & Xiao, 2006; Gittell, 2002). Unstructured coordination describes how agents coordinate ad hoc and in uncertain situations in the absence of pre-specifications (Argote, 1982), structure, and formality. In unstructured coordination, agents need to deal with the "'situation' [rather than] … formal organizational arrangements" (Faraj & Xiao, 2006, p. 1157). In unstructured coordination scenarios, agents face uncertainty because they have little information about expected interactions (Gittell, 2002), which makes it difficult to predetermine coordination activities (Argote, 1982). Agency is thus important for unstructured coordination (Claggett & Karahanna, 2018, p. 709). In the case of human-AI ecologies, both human and AI-based agencies matter (Baird & Maruping, 2021). To understand coordination among human and AI-based agents, we draw from research on human-computer interaction (HCI) and multi-agent systems (MAS).

Coordination and Human-Computer Interaction

Integrating research from computer science, psychology, and social sciences, HCI studies the relationship between human and AI-based agents, typically with an eye toward improving fit under varying circumstances (Carroll, 1997). HCI conceptualizes both human and AI-based agents as information processing systems (Carroll, 1997; Limerick et al., 2014). Both types of agents have the ability to sense, interpret, and use information to enact certain outcomes. The information processing of both systems can mutually influence each other (Harrison et al., 2007).

HCI research increasingly acknowledges that interactions between human and AI-based agents can be situated in complex, unpredictable environments (Frauenberger, 2019; Harrison et al., 2011). AI-based agents pose particular challenges due to their unpredictability (Limerick et al., 2014). Because such systems can act autonomously and learn over time, it is difficult to determine which agent is responsible for what outcome or to clearly pinpoint how technology can serve human users more effectively (Frauenberger, 2019). Complex, real-world situations characterized by uncertainty require actors' attention to abstract issues, such as ethics, trust, and relationships (Kuutti & Bannon, 2014). HCI explores situated configurations of human and AI-based agents that both influence each other and give rise to dynamically changing, even unexpected, relationships over time (Frauenberger, 2019; Harrison et al., 2011).

Building on these observations, HCI research views interactions between humans and AI-based agents as dynamically unfolding on the fly (Harrison et al., 2007) and in the wild (Browne & Garnham, 2022). HCI researchers are increasingly theorizing about interactions to understand change and learning in ways that can be applied to make sense of specific situations (Frauenberger, 2019). Predictive processing theory is one such theoretical perspective to help explain how human and AI-based agents make sense of and interact with their environment under different levels of certainty (Browne & Garnham, 2022). Predictive processing theory is an emerging theory in computational neuroscience that describes information processing for both human and AI-based agents (Clark, 2013; Hohwy, 2013). According to this view, the human brain and learning algorithms of AI-based agents are both processors that project predictions from knowledge accumulated in the past to understand incoming sensory stimuli.

Coordination and Multi-Agent Systems

Research on multi-agent systems (MAS) investigates how multiple software-based agents[2] can be designed to work together in a task environment to solve a given problem (Stone & Veloso, 2000). A key question in this stream of research is how coordination and control can be ensured, especially when agents are heterogeneous, have limited means of communication and negotiation, and require complex and dynamic problem-solving (de Freitas & dos Santos Silva, 2017; Stone & Veloso, 2000). In such contexts, heterogeneous agents can have complementary skills and features, may be responsible for carrying out different aspects of a task at different points in time, and may interact in collaborative but also competitive ways (Frantz & Pigozzi, 2018; Stone & Veloso, 2000). Thus, MAS research has explored, to some extent, how agents coordinate in complex, unstructured, and uncertain situations.

[2] We acknowledge that the multi-agent systems literature also considers the role of humans as agents (e.g., Stone & Veloso, 2000). However, in the context of norms, the emphasis is typically on collectives of non-human, machine-based agents and their interactions.

One approach in dealing with uncertainty and unstructured situations in MAS emphasizes flexible coordination through social conventions or norms (Stone & Veloso, 2000; Walker & Wooldridge, 1995). Norms are restricted sets of action templates available to agents in situations that require coordination (Walker & Wooldridge, 1995). Norms refer to shared expectations among agents about what should happen in a given situation without requiring explicit communication between agents (Stone & Veloso, 2000). While norms are similar to rules in that they prescribe appropriate agent behavior, the key difference is that norms prescribe general patterns of social behavior that can be less fixed and more discretionary (Savarimuthu & Cranefield, 2011). Explicit and strongly reinforced norms are essentially rules, whereas vague and weakly reinforced norms are more like suggestions. Norms allow agents some flexibility in terms of how they act and interact (e.g., Frantz & Pigozzi, 2018; Stone & Veloso, 2000).

Norms entail an obligation, request, or prohibition with respect to available action opportunities, often bearing ethical or moral implications (Chopra et al., 2018). Existing work offers recommendations for what should or should not be part of norms in MAS, including issues around the emergence of norms (e.g., Frantz & Pigozzi, 2018). Collectives of AI-based agents develop interaction patterns through pre-specified norms, or they develop norms endogenously as a growing share of agents enact an emergent interaction pattern (Frantz & Pigozzi, 2018).

Taken together, the fields of HCI and MAS offer complementary views and emphasize different aspects of coordination relevant to human-AI ecologies. HCI research stresses the role of humans and how their information processing activities are key to understanding interactions with AI-based agents. Research on MAS is concerned with technical applications to address unstructured coordination between AI-based agents. While the coordination among AI-based agents cannot be fully prescribed, it can be enabled through the emergence of norms that instill stability in the system while providing freedom for situated interactions in response to specific contexts. Figure 1 provides an overview of these two research streams and how they are synthesized for the purpose of this article.

Kernel Theories for the Management of Human-AI Ecologies

Ensuring effective and ethical activity within human-AI ecologies is perhaps the key management challenge of our time. This is because AI-based agents are dramatically increasing in performance and broadening in scope (Berente et al., 2021). Their potential to transform social institutions and produce large-scale social outcomes is becoming widely apparent (Kissinger et al., 2021; Rahwan et al., 2019), and this transformation needs to be managed (Kane et al., 2021). Therefore, we propose a design theory as a guide for the management of human-AI ecologies. Rooted in concepts from HCI and MAS, we infer three central aspects that are relevant for the conceptual development of our design theory:

1. Coordination is an information processing activity that involves perception, interpretation, and action (Faraj & Xiao, 2006; Gittell, 2002). A design theory for norm-based coordination in human-AI ecologies needs to account for information processing in all types of agents involved in interaction and, hence, coordination in human-AI ecologies.

2. Coordination emerges and changes. A design theory for norm-based coordination in human-AI ecologies must balance needs for prespecified rules with the ability to respond to changing features and situations. One way of thinking about this, according to MAS research, involves the emergence of norms (Frantz & Pigozzi, 2018).

3. A design theory for norm-based coordination in human-AI ecologies must be formulated in a way that can be applied to different types of AI-based agents. For example, agents can have different features and can be based on supervised, unsupervised, or reinforcement-based learning (see, e.g., Weiss, 1999).

We draw on two kernel theories to lay the foundation for a design theory for norm-based coordination in human-AI ecologies: predictive processing theory as a unifying lens for information processing in human and AI-based agents and social norm theory to explain coordination through shared expectations about each other's behavior (Hawkins et al., 2019).

Kernel Theory I: Predictive Processing

Predictive processing theory is an emerging theory in the cognitive sciences and computational neurosciences that views the brain as a Bayesian prediction machine (Clark, 2013; Hohwy, 2013). The brain is regarded as a proactive organ that is concerned with projecting knowledge accumulated in the past onto incoming sensory stimuli. Predictive processing theory is considered a paradigm shift in the understanding of cognition, action, and perception (Friston et al., 2017).

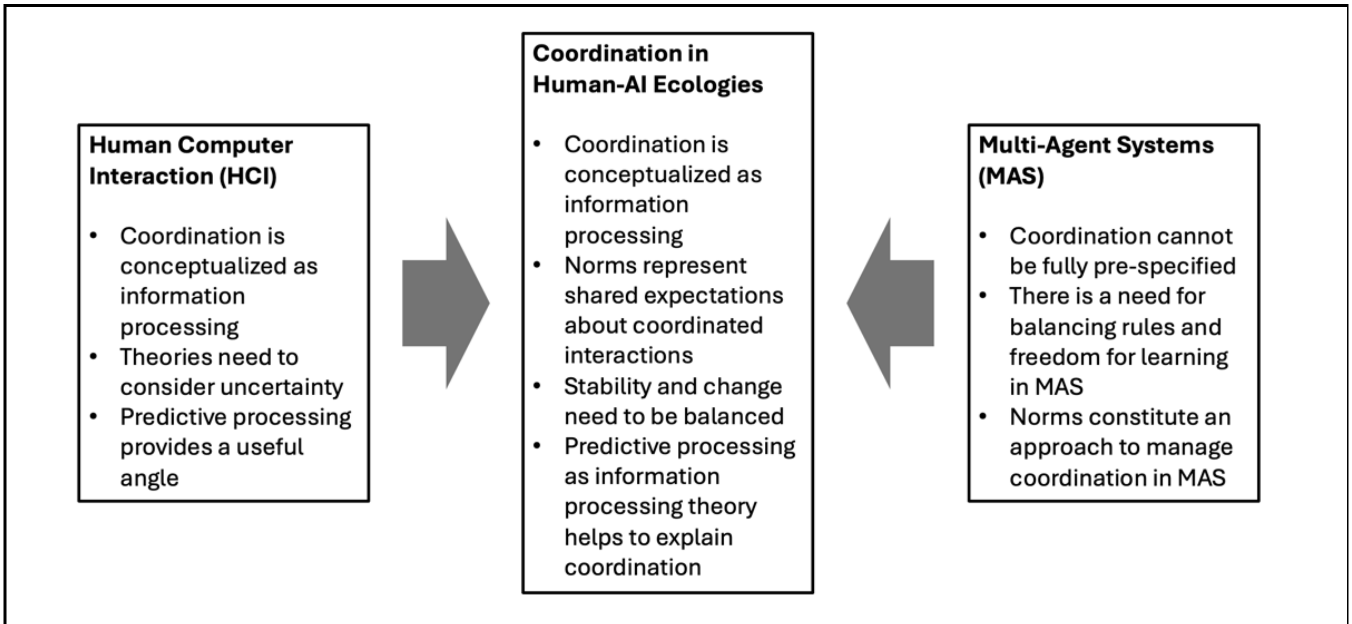


Figure 1. Synthesis of the HCI and MAS Literature to Understand Coordination in Human-AI Ecologies

Predictive processing theory provides an overarching framework to describe information processing principles for both human *and* AI-based agents (Ciria et al., 2021; Clark, 2016; Lanillos & Cheng, 2018; Taniguchi et al., 2023). It presents central mechanisms that explain human-related cognitive phenomena, such as mental dysfunctions (Hohwy, 2013), as well as AI-related information processing aspects, such as the development of robotic movements (Ciria et al., 2021). The lens allows us to describe information processing for various kinds of machine learning applications. Regardless of whether machine learning is based on supervised, unsupervised, or reinforcement learning, predictive processing highlights how artificial neural networks generate predictions at different scales. Positive experiences strengthen predictions and increase their precision in the future, whereas unsuccessful predictions provide prediction errors that might indicate a need to adjust the underlying model and the resulting output (Russell & Norvig, 2020).

Predictive processing theory further acknowledges that agents perceive, act, and learn *in relation to* other agents (Clark, 2013). They develop predictions that consider their own action repertoires and those of others (Kelly et al., 2019). Predictive processing is based on the assumption that agents seek to reduce uncertainty in coordination by establishing routinized social practices, such as norms, to provide stable and predictable behaviors (Clark, 2018; Williams, 2018). In what follows, we outline the basic principles and mechanisms of predictive processing theory to apply this lens to coordination in human-AI ecologies.

Basic Principles of Predictive Processing

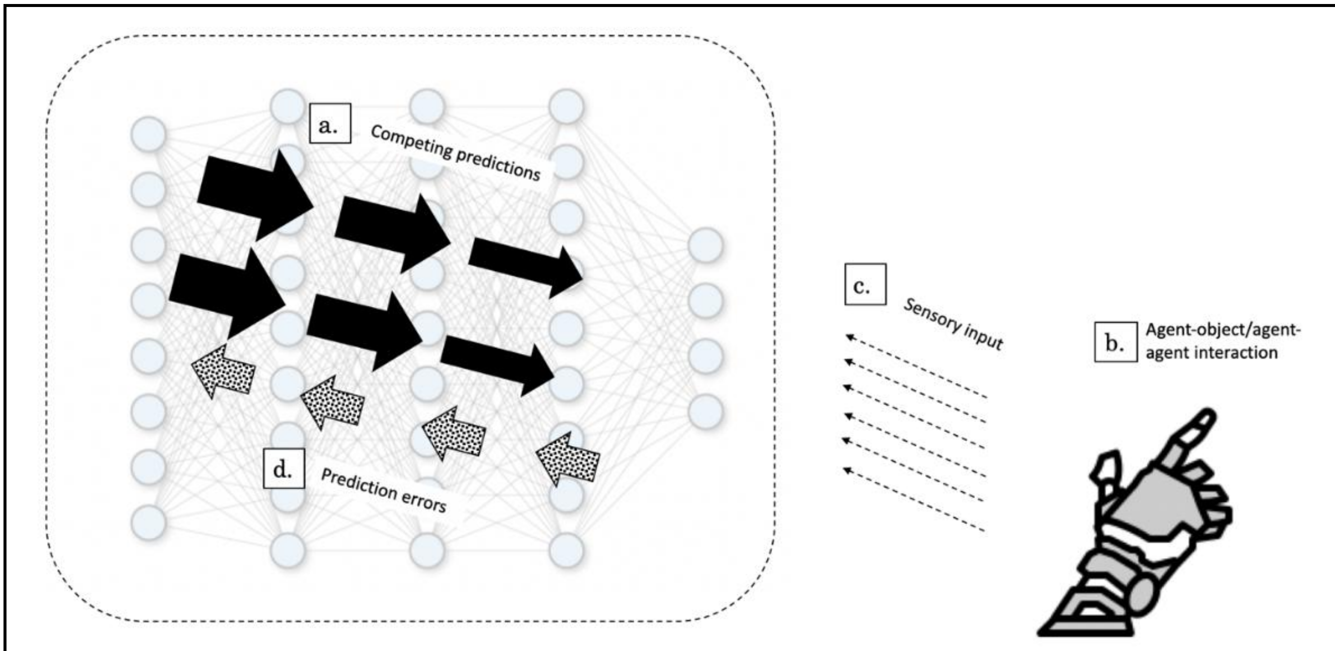
Predictive processing theory holds that human and AI-based agents are experience-driven prediction machines that constantly evaluate the dynamics in their external environment to make inferences about what is most likely to happen next (Ciria et al., 2021; Linson et al., 2018). Predictive processing works as follows (e.g., Clark, 2016). Prior knowledge about a given environment is translated into predictive models, which contain predictions about what is most likely to happen in a specific situation. Predictions in such models are nested and occur at different hierarchical levels, dependent on the complexity of an agent's internal model. Higher-level predictions are increasingly abstract and entail broad expectations that are projected onto and provide prior knowledge for lower-level predictions. Lower-level predictions become more granular, ultimately translating into sets of expectations about specific incoming sensory signals. Thus, top-down predictions translate across the hierarchy from higher to lower levels. Incoming sensory signals are tested against these top-down predictions. They either align with top-down predictions or not, and this impacts how predictions evolve over time. The mechanism that explains this evolution is based on the concept of *prediction error* (Clark, 2016).

Prediction errors occur when incoming sensory signals are surprising and challenge established predictions (Clark, 2013, 2016; Linson et al., 2018). As agents evaluate predictions on the grounds of the expected outcome and the actual outcome of an action (Reichardt et al., 2020), prediction errors result from differences between the expected sensory signal and the actual incoming sensory signal. They represent the degree of uncertainty that an agent is confronted with at a given point in time. If an agent is not confronted with any prediction errors, the predictive models are accurate, and the agent can further carry out tasks and actions in expected ways (Linson et al., 2018). In this situation, the agent does not update its models. If, however, the agent encounters a prediction error, this error may "call for an update" (Den Ouden et al., 2012, p. 2). Hence, there may be learning to refine a predictive model in such a way that it reduces prediction errors in future encounters (Reichardt et al., 2020).

The assessment of prediction errors is reflected in the models' precision weights (e.g., Ciria et al., 2021; Clark, 2016, 2018). Precision weights are focused estimates of an agent's own certainty about incoming sensory signals. While predictions draw on prior knowledge to encode relevant states of the environment, precision weights represent the agent's own certainty of how trustworthy these incoming sensory signals are. Precision weighting—the process of assigning precision weights to sensory signals and prediction errors—is the process of extracting relevant signals from noise. If an agent carries out tasks while entertaining predictions with high expected precision, it will assign high precision weights to specific sensory signals and likely account for prediction errors (i.e., deviations are not expected and thus create significant errors).

The agent is continually engaged in attempts to predict its own context-dependent certainty of incoming sensory signals. There is not a singular model, prediction, or prediction error, but an ongoing stream of multiple, hierarchical, and competing predictions. The upshot of prediction competition is that an agent depicts the world as an "evolving matrix of parallel, partially computed possibilities for action and intervention" (Clark, 2016, p. 251) to select a prediction that seems most appropriate in a given situation.

Context is at the core of prediction competition. Agents continuously compute alternative predictions based on different probabilistic models that might be relevant to recognize a given situation and take appropriate actions that have the best fit (Clark, 2016). When contextual factors change, agents' predictions change as well because "[w]hat is in the foreground and what is in the background … constantly shifts" (Kiverstein & Rietveld, 2012). Figure 2 summarizes the basic mechanisms of predictive processing theory.



Note: (a) An agent—human or AI-based—is equipped with a neural network, i.e., its internal model (within the dashed line). The internal model generates predictions about the environment. (b) The agent acts (represented through the robot hand) in correspondence to its predictions. (c) The action generates feedback in terms of sensory signals. (d) Sensory signals that do not correspond to predictions turn into prediction errors; they are passed on to higher levels and are either ignored or change the agent's internal model. The relevance of prediction errors is determined through precision weights.

Figure 2. Schematic Representation of the Basic Mechanisms of Predictive Processing

Coordination Through Predictive Processing

Predictive processing theory is applicable to settings where multiple agents interact (Clark, 2016; Kelly et al., 2019; Tamir & Thornton, 2018; Williams, 2018) and is thus also applicable to human-AI ecologies. Agents interact with an environment that is populated with other agents. Successful interaction involves coordination, which in turn depends on an agent's ability to predict other agents' behaviors (Hawkins et al., 2019; Tamir & Thornton, 2018) and requires knowledge about unobservable states and processes in those other agents, such as intentions, as well as about observable actions and their consequences (Clark, 2016). Coordination within social contexts depends on shared practices to "both predict and be predicted by others" (Wheeler et al., 2020, p. 192). Agents incorporate the behavior of other agents into their own models to reduce processing loads. An agent might even constrain its own behavior to be more predictable for other agents (Clark, 2016). Coordination requires shared predictive models that ensure "socially constructed normativity" against which agents assess the validity of their individual models (Williams, 2018, p. 855). This normativity requires some measure of shared knowledge in the form of language, routines, or rules for interaction. Through norms, agents reduce uncertainty and enable coordination.

Kernel Theory II: Norms for Coordination

Norms provide the predictability that allows agents to coordinate effectively. Norms are social conventions that are formed and shared among agents and provide shared expectations and stable interaction patterns (Feldman, 1984; Hawkins et al., 2019; Lewis, 1969; Marmor, 2009). There are a variety of perspectives about norms in the literature (e.g., Horne & Mollborn, 2020) that highlight how norms have a rule-like quality for establishing and maintaining order and meaning in social situations (e.g., Chung & Rimal, 2016; Horne & Mollborn, 2020). We focus on understanding norms *for coordination*, that is, norms that involve some kind of empirical expectations about how others behave in a given situation but do not necessarily carry any moral or ethical implications (e.g., Feldman, 1984; Hawkins et al., 2019; Ullman-Margalit, 1977).

Norms for coordination develop as multiple agents learn about the most probable interaction in a given situation. They emerge through social interactions in recurrent and comparable situations (Hawkins et al., 2019; Ullman-Margalit, 1977). Because norms develop through the repetition of certain coordinating actions, they involve collective awareness and widespread expectations about shared behavioral patterns for recurrent coordination situations that others expect or approve of (Chung & Rimal, 2016; Horne & Mollborn, 2020). To the extent that norms are enforced to simplify and make the actions of others predictable, they constitute a relatively persistent and binding condition for interactions (Feldman, 1984), primarily because agents become increasingly familiar with them and thus expect that they "*should* be carried out" (Przepiorka et al., 2022, p. 2). Norms can also change. Over time, norms may become out of date and misaligned with social imperatives, which can be problematic for the social context. When certain conditions inside or outside a social collective threaten the stability of a norm, the norm will likely be adjusted (Hawkins et al., 2019; Marmor, 2009). To understand how norms change over time, it is useful to think of different manifestations of norms: some norms are general and abstract (deep norms), whereas others are specific and situated (surface norms).

Types of Norms

We distinguish between two types of norms: *deep norms* and *surface norms*.[3] Deep norms are general and abstract rules that apply across contexts and social situations and can often be non-conscious (Horne & Mollborn, 2020). Surface norms imply more specific yet also flexible behavioral rules that can change in response to changing requirements for coordinated interactions.

Deep norms: Deep norms are persistent and "sticky" in a social collective and have considerable stability over time (Przepiorka et al., 2022). They are abstract conventions—generally appropriate standard social behaviors—that permeate across contexts. Agents may not be explicitly aware of such underlying norms, which provide an unconscious and taken-for-granted frame of reference for interactions (Horne & Mollborn, 2020). Deep norms relate to general expectations about how agents should act in everyday situations, such as being polite and respectful when they encounter and greet one another for the first time.

Deep norms provide the foundation for the emergence of more specific norms, which are then enacted by subgroups of the larger collective or in a smaller set of contexts and situations. Deep norms represent a shared expectation among all agents

[3] We borrow the terms "deep" and "surface" from Marmor's (2009) notions of deep and surface conventions. Marmor uses the term "convention," which is consistent with what we mean by norms, namely shared behavioral expectations among agents to accomplish coordination. Marmor argues that conventions differ from norms because the former emerge in arbitrary ways. In line with research on norms, particularly "coordination norms" (Horne & Mollborn, 2020; Ullman-Margalit, 1977), we do not make this distinction.

(Horne & Mollborn, 2020) and provide the "broader population-level infrastructure of existing norms" in which interactions in new situations are embedded (Hawkins et al., 2019, p. 167). Deep norms correspond to what research on multi-agent systems refers to as the offline design of norms, which are inscribed as top-down specifications into an agent's behavioral repertoire prior to its interaction with other agents (Morales et al., 2013; Savarimuthu & Cranefield, 2011; Walker & Wooldridge, 1995).

In general, the more heterogeneous the social collective, the less specific shared expectations about interactions become, with deep norms reflecting increasingly generalized guidance (e.g., Abrutyn & Carter, 2015). For instance, a deep norm of courtesy generally applies as a greeting, yet specific norms may vary across contexts and groups, such as a handshake in the West or a bow in the East. Such specific norms are surface norms.

Surface norms: What we refer to as *surface norms* instantiate deep norms as specific conventions or rules for action in specific situations that incorporate what is specified in deep norms. Surface norms emerge in new and unknown situations when a collective of agents seeks to establish expectations about appropriate interactions. If the same or similar agents in a homogeneous population face recurrent coordination scenarios, surface norms will remain relatively stable, and little innovative normative behavior will emerge. When social collectives face new situations, however, different surface norms are more likely to emerge. Heterogeneous and pluralistic collectives will likely develop a variety of surface norms to account for a larger variety of behaviors among agents (Abrutyn & Carter, 2015). This pluralism implies less stability of surface norms, since new norms continuously emerge and, in some cases, replace old norms. Consider, for instance, how etiquette manuals in the United States have changed across editions in the sense that normative recommendations become more abstract, offering "general guidelines" for appropriate behavior, while "the details and differences in the context" became subject to continuous adaptations in specific situations and subgroups (Abrutyn & Carter, 2015, p. 361). The enactment of specific norms occurs in subgroups and can change quickly. From this perspective, surface norms correspond to what research on multi-agent systems refers to as the "online" design of norms, where new norms emerge on the grounds of deep norms as agents engage in coordinated interactions (Morales et al., 2013; Savarimuthu & Cranefield, 2011; Walker & Wooldridge, 1995).

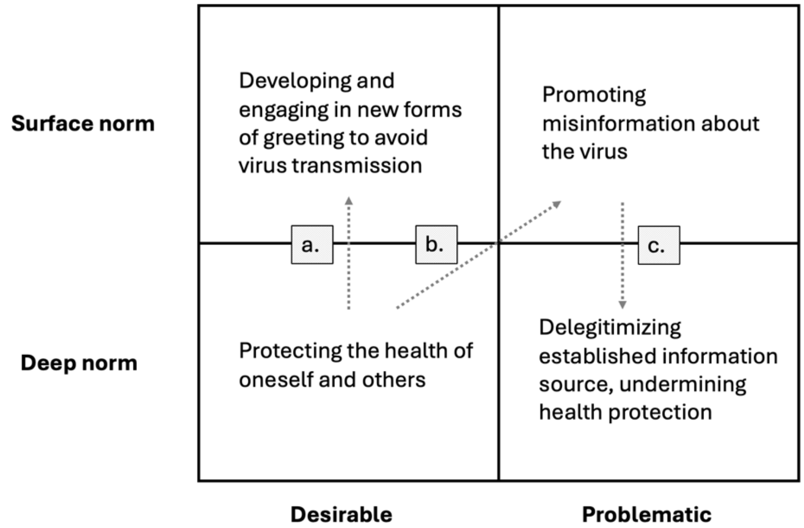
Interplay Between Deep Norms and Surface Norms

Deep norms and surface norms influence each other in different ways. These can be desired, such as when collectives of agents are exposed to potentially harmful events. An illustrative context is the COVID-19 pandemic, where norms emerged to account for situational constraints, such as new forms of greeting that reduced the risk of the virus spreading (Diekmann, 2022). In this example, deep norms involving polite greetings in conjunction with deep norms about the importance of protecting health (Iacono et al., 2021) led to the emergence of surface norms that reflected expectations about how to coordinate behavior (Diekmann, 2022).

The relationship between deep and surface norms is not always positive and can lead to adverse consequences. Norms can also be "bad," especially when they emerge as problematic or harmful instantiations of deep norms (Abbink et al., 2017). The same deep norm can lead to the emergence of surface norms that raise ethical concerns. Consider, for instance, how collectives of agents start to disseminate misinformation about a virus from dubious news sources (Loomba et al., 2021), all the while pursuing the deep norm to protect health. In other words, it is crucial to understand what kinds of surface norms emerge on the basis of deep norms and, vice versa, what kind of surface norms are reinforced and firmly embedded in a social collective, ultimately taking on the form of deep norms. For example, the deep norm of protecting health may translate into new greeting rituals involving bows, fist-bumps, elbow bumps, or nods rather than handshakes or hugs, and thus new surface norms may emerge that reinforce those deep norms. New surface norms may also contribute to changing deep norms. For example, one may translate the deep norm of protecting health to the dissemination of misinformation, which may cause agents to develop general distrust in established media sources. Over time, agents may act in ways that ignore or delegitimize official information sources, and their distrust towards official media sources may be reflected in a deep norm that actually undermines the original deep norm. Figure 3 illustrates the relationship between deep and surface norms.

Integrating Predictive Processing and Norm Theory

From the perspective of predictive processing theory, norms are shared predictive models about the behavior of other agents that facilitate coordination to the extent that other agents share the same model (see Wheeler et al., 2020). As such, norms represent "mutually constraining error minimization processes" that constrain the range of interactions in social contexts (Wheeler et al., 2020, p. 192). Deep norms give rise to surface norms; surface norms inform coordinated interactions and change in response to prediction errors, which may lead to changes in deep norms (see Figure 4).



Note: Socially desirable deep norms can lead to (a) desirable surface norms and (b) problematic surface norms. Problematic surface norms may manifest as (c) problematic deep norms.

Figure 3. The Relationship Between Deep Norms and Surface Norms

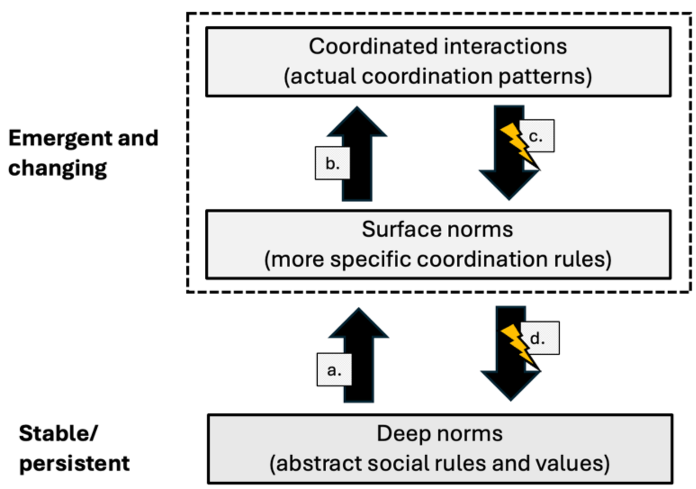


Figure 4. The Relationship Between Deep Norms, Surface Norms, and Coordinated Interactions in Human-AI Ecologies

Deep Norms Reflect Basic Values and Give Rise to Surface Norms

Deep norms (lower grey bar in Figure 4) represent foundational rules and values for coordination. From the viewpoint of predictive processing theory, deep norms are predictions with expected high precision in relation to relevant contextual cues. Specific expected sensory signals that correspond to cues in the environment are assigned high precision weights because the predictive model expects them to be present in a situation. If they do not occur as expected, high prediction errors occur.

Deep norms give rise to the development of surface norms (Figure 4, arrow "a"). Surface norms instantiate deep norms and emerge in response to specific situational coordination needs. Surface norms, along with specific coordination needs, are emergent and can dynamically change (Kelly et al., 2019) (see "online" design of norms in the multi-agent systems literature).

Surface Norms Guide the Enactment of Coordinated Interactions

Surface norms inform specific interactions among agents in coordination scenarios (Figure 4, arrow "b"). Coordinated interactions occur in specific situations where two or more agents enact shared predictions about each other's expected actions. The specific surface norm on which a coordinated interaction should be based depends on the context of the coordination scenario. When a coordination scenario is known (e.g., because agents have encountered it before), agents can select a prediction representing a surface norm with expected high precision among competing predictions by considering contextual cues (e.g., based on how another agent behaves and what it signals to the agent). In such a coordination scenario, the surface norm will lead to the enactment of a low-level prediction shared by all involved agents. The agent encounters low prediction error because it can predict the sensory signals that result from coordinated interactions. However, it can also be the case that the coordinated interaction cannot be predicted based on existing surface norms, such as when the coordination scenario is unknown.

Uncertainty in Coordination Scenario Leads to Prediction Error in Surface Norm

When agents encounter uncertainty in coordination scenarios, they cannot rely on established predictions about how they should interact. Uncertainty is represented through prediction errors (Figure 4, arrow "c"; the prediction error is depicted as the yellow flash symbol). Prediction errors occur when predictions are not reliably accurate. That is, incoming sensory signals from actual interactions in the environment do not match the expected sensory signals that are built on the grounds of existing surface norms (Clark, 2016). When a situation is often repeated, there is typically expected high precision, but when actual interactions vary from this expectation, the agent encounters prediction errors.

Incoming prediction errors may indicate a need for the agent to adjust an existing surface norm or develop a new one to minimize prediction errors in future coordination scenarios by adjusting its hierarchical prediction models. These updates need to incorporate the agent's expectations about the situation as well as expectations about what other agents are likely to do. A new surface norm will account for preferences of human or AI-based agents to varying degrees.

Importantly, there are at least two cases (see, e.g., Den Ouden et al., 2012) where prediction errors should not always lead to an update of predictions about surface norms. First, when a given prediction error can be explained away using simple adjustments at low levels, such as adjusting physical movements during the interaction, it may not require a change to a surface norm or the development of a new one. Second, if the prediction error is associated with high-level predictions and associated contextual cues, such as when it violates a deep norm, it should often avoid updating the model because the situation may reflect an anomaly, and updating the model would undermine future coordinated interactions and threaten the values of the human-AI ecology.

Prediction Errors in Surface Norms Can Lead to Prediction Errors in Deep Norms

When surface norms are considered reliable and effective and do not violate deep norms, they can evolve into deep norms. This is the case, for example, when human developers identify that a surface norm reflects accepted values and translate it into "hard-wired" high-level predictions. This can decrease uncertainty in certain situations and encourage desirable behavioral patterns across the ecosystem. As they are enacted and adjusted, these deep norms are reinforced and strengthened. Subsequent surface norms formed on the grounds of the new deep norm will further embody the expectations captured in the deep norms (Figure 4, arrow "a").

Further, it is important to note that agents may not only encounter prediction errors about surface norms but also deep norms (Figure 4, arrow "d"). Deep norms are high-level predictions with high expected precision in relation to specific contextual cues. When agents' interactions generate large prediction errors with respect to specific predicted sensory signals, they possibly violate a deep norm. In such cases, it is generally appropriate for agents to avoid changing their surface norms.

Illustrative Example: Autonomous Warehouse Ecologies

Settings where human-AI ecologies are becoming increasingly prevalent abound: self-driving vehicles, virtual game environments, autonomous robotic production and operations, bots and humans interacting in online communities, a variety of AI agents used in workflows, and others. To illustrate our argument, we draw on a context where autonomous agents and humans have been interacting for years and are expected to increasingly do so in the future: autonomous warehouses (e.g., Morris, 2024; Shaheen, 2023). Autonomous warehouses are inhabited by human workers and different types of AI-based agents. These agents have complementary roles and skills and are responsible for a variety of tasks. In such warehouse settings, “humans have to adapt to the machines as much as the machines have to adapt to us” (Simon, 2019). For instance, Amazon reports using AI-based agents in its warehouses, which operate alongside humans by handling packages, managing inventories, and delivering products (Quinlivan, 2023).

Such autonomous warehouses are an early instance of human-AI ecologies (from hereon: “warehouse ecologies”). They are environments where human agents and AI-based agents share the same cues and resources and operate on centralized and, conceivably in the future, decentralized control. Furthermore, human and AI-based agents learn from one another over time and develop coordinated interactions. Whereas current warehouse ecologies are based on the assumption that all involved AI-based agents are designed to function within a more or less specific context or system, we can conceive of future scenarios where a variety of agents will interact, coming from different vendors and working across different boundaries.

Warehouse ecologies are characterized by goals and tasks that bind all agents and remain stable over time. At the same time, agents need to flexibly adjust to unexpected situations (Blake, 2023). Hence, the context of warehouse ecologies allows us to theorize about planned and emergent coordination in human-AI ecologies based on existing reports, as conveyed, for instance, in public media (Blake, 2023; Schreiber, 2019), while also accounting for value-based considerations—for example, in relation to the treatment of human workers.

The safety of human workers is considered of utmost importance in the design of AI-based agents that interact in warehouse ecologies (Johnson, 2022). However, the interplay between AI-based agents and humans can lead to various problems, ranging from coordination problems to situations where humans are subject to safety violations and even physical harm (Johnson, 2022). This is also because warehouse ecologies yield different sources of uncertainty, such as blind corners, intersections between human and autonomous drivers, and poor flooring (Blake, 2023), or new orders that require varying, potentially more complex handling (Simon, 2019). Since it is impossible to foresee every kind of interaction, agents in warehouse ecologies can face unstructured coordination scenarios, requiring them to adjust their behaviors and collectively develop new norms.

Deep norms may involve general values—for example, that robots should never harm humans—and very basic rules that need to be adhered to by all agents, such as that one has to stop in front of another robot or obstacle (Johnson, 2022). Deep norms pattern how agents can learn and change their interactions in human-AI ecologies. Ideally, the same deep norms will apply to all agents, meaning that all agents will act in accordance with the same values. Therefore, the design of deep norms is essential for human-AI ecologies, and managers must know the relevant values that underly all interactions, reflect on what they mean for the ecology they manage, and be able to embed them in the ecology. In warehouse ecologies, this involves complex decisions—for instance, regarding the resolution of tensions between efficiency and safety. Certainly, the intuitive tendency may be to emphasize safety, but, in reality, not every situation that is potentially harmful for a human can be anticipated. Also, organizations must pursue efficiency to remain competitive, and managers will only be willing to compromise on it to a certain extent. Since the design of deep norms in warehouse ecologies can be a complicated process, deep norms may be modified and adjusted in response to specific outcomes over time. For example, just as definitions of fairness may change over time due to situations evolving (Berente et al., 2024), one might expect different deep norms for efficiency and safety in autonomous warehouses to evolve over time.

The specific outcomes of agent interactions in warehouse ecologies are based on surface norms. For example, when deep norms require safety—in the context of safety laws, for example—they will be instantiated in surface norms to guide specific coordinated interactions. The deep norm of safety may be instantiated with the help of certain zones for motion in the warehouse. This enactment may then result in task-specific features and shared cues in the warehouse environment (Blake, 2023), such as yellow border lines and cameras that can recognize these lines. Managers may implement surface norms into task-specific features by hard-coding rules about where agents can travel in a warehouse or by training agents using simulations, for example. For a warehouse ecology to evolve and change over time, agents should also be able to develop new surface norms on their own and without the intervention of managers. For instance, when they encounter uncertain coordination scenarios, they may develop new interactions. Deep norms must be taken into account when agents update their prediction models to implement new norms. Consider, for example, how an AI-based agent in a warehouse ecology observes that a certain process is carried out much more efficiently while, at the same time, a human is put at risk, for example, by a falling item. Such a situation should create a high prediction error when the deep norm emphasizes safety. Despite the apparent efficiency gains, the autonomous agent must therefore not update its internal model for future operations. If such incidents happen repeatedly, agents may adjust surface norms and wait patiently for the human worker to fulfill the task while taking all required safety measures (see Evans, 2020). Tasks are accomplished by adequately dealing with trade-offs such as efficiency versus safety, which is critical for warehouse ecologies to be effective (Evans, 2020).

Design Principles for Norm-Based Coordination in Human-AI Ecologies ■

In this section, we present a design theory based on a set of interrelated design principles for norm-based coordination in human-AI ecologies, using the warehouse ecology example for illustrative purposes.

Assumptions About Roles

Our design theory distinguishes between three roles. It seeks to inform *managers* and *developers* who collaborate to set up and continuously manage AI ecologies involving *agents* (Table 1). *Managers* determine the core values and rules of human-AI ecologies. They are decision makers who may act on behalf of companies, governmental bodies, or other organizations. Their main interest is that the human-AI ecology performs within desired boundaries (Gasser & Mayer-Schönberger, 2024). Human-AI ecologies involve heterogeneous *agents*, both human and AI-based, and are extensible. It may not always be possible for managers to influence all agents. *Developers* have the skills and competences to establish and then continuously modify human-AI ecologies by providing means for managers to (1) influence agents by encoding instructions into AI-based agents, such as through training, and (2) monitor both human and AI-based agents and their interactions.

In human-AI ecologies, development and management are two sides of the same coin, since intervening in a human-AI ecology may require changing the hierarchical prediction models of involved agents, which may involve training, which is a key design activity (Russell & Norvig, 2020). In some cases, the same individual may assume the role of only a manager, a developer, or both; for analytical purposes, we discuss the two roles separately. In the setup stage of a human-AI ecology, managers make assumptions about the environment in which agents interact with other agents. They decide on the tasks and goals of agents, along with the associated features and skills, and, with the help of developers, they train agents so that they interact in desired ways. This involves reflections on norms. Once a human-AI ecology is in place, it must be continuously monitored and adjusted. Managers must assess how the ecology is performing and need to be able to adjust the normative interactions of some or all agents, which might require development capabilities. Developers are therefore involved before agents are introduced in human-AI ecologies and continue to modify and adjust those agents over time.

Design Principles

Human-AI ecologies require shared mutual predictions that facilitate as well as constrain the interactions between agents, both human and AI-based. On the one hand, shared mutual predictions must be maintained to continuously prevent the emergence of undesired norms and outcomes. On the other hand, agents must be capable of and allowed to generate new coordinated interactions, which may become new norms in response to unknown and changing coordination scenarios. Deep norms guide the emergence of new surface norms, and surface norms are encoded either through learning or discrete commands to be enacted by agents. Managers instill values and rules by making decisions and communicating shared norms to provide the stability that allows for reliable coordination and the flexibility to maintain coordination in the face of unpredictability.

Before we proceed with the development of the design principles, we provide an overview of key concepts in Table 2 based on social norm theory and predictive processing theory. These are important building blocks for defining the mechanisms (e.g., implementing values as high-level predictions with high precision) underlying the rationales (e.g., high prediction errors in predictive processes) and intended outcomes (e.g., shared norms) of our design principles (Gregor et al., 2020).

Table 1. Roles in Human-AI Ecologies

Role	Description
Managers	Decide on core values and basic rules; managers act on behalf of organizations or institutions and can have an influence on all or some agents in the human-AI ecology.
Developers	Establish and modify the setup of human-AI ecologies by providing means for encoding instructions in agents' internal models.
Agents	Act within human-AI ecologies and can be human or AI-based; agents interact with one another as they pursue individual and shared goals.

Table 2. Key Concepts for Design Principles, Synthesizing Predictive Processing Theory and Norm Theory

Concept	Explanation
Deep norm	High-level, general shared prediction with high precision weights assigned to sensory signals that reflect relevant contextual cues in a coordination scenario.
Surface norm	Low-level, specific shared prediction providing guidance for coordinated interactions; instantiating deep norms and assigning precision weights to sensory signals representing additional contextual cues in a coordination scenario.
Coordinated interaction	Specific interaction in a human-AI ecology where two or more agents enact shared predictions (represented as deep and surface norms) about mutually expected interactions.
Expected precision of prediction	Agent's own certainty about the accuracy of a prediction; determines how reliable a prediction is with regard to incoming sensory signals, e.g., in response to an interaction; in predictions with high expected precision, specific sensory signals are assigned high precision weights.
Contextual cue	Salient cues in a coordination scenario that represent the relevant context; they are relevant for deep and surface norms when conveyed as sensory signals with high precision weights and can be present in the situation or transmitted by agents.
Low-uncertainty coordination scenario	Specific situation where agents engage in coordinated interaction and where they can draw from existing predictions (i.e., norms) with expected high precision (e.g., because they have repeatedly faced a specific coordination scenario).
High-uncertainty coordination scenario	Specific situations where agents engage in coordinated interaction, but the existing predictions have low expected precision (e.g., because the agent has not faced this situation before).

While we emphasize that both types of agents—human and AI-based—are part of human-AI ecologies, the focus of our design principles is on the design of AI-based agents. This is based on our assertion that both types of agents enact norms and develop new behaviors based on the prediction errors they receive from their interactions with other agents. The assumption is that both can be understood as Bayesian prediction machines that use previous experiences to predict what is likely to happen next: They change and adjust their predictions when they encounter prediction errors. While human agents cannot be "designed," they can be informed about deep norms and surface norms and, of course, play a key role in creating known and unknown coordination scenarios in human-AI ecologies. For each design principle, therefore, we highlight the intended design outcome: coordinated interactions among human and AI-based agents based on the assertion of symmetry in terms of how these agents enact surface norms consistent with deep norms to reduce uncertainty and how they collectively develop new norms.

We formulate four design principles that fall into two broad categories representing two aspects involved in managing human-AI ecologies: *managerial intervention* (Table 3) and *agent interaction* (Table 4). For each design principle, we provide more detailed subprinciples to highlight the intended context, outcomes, and underlying mechanisms (Gregor et al., 2020). To this end, we formulate all specific principles as either "managers should be able to…" or "an agent should be able to…" (with slight variations). The recipients of these principles are managers collaborating with developers in setting up and operating a human-AI ecology. They are the "implementers" of the design principles (Gregor et al., 2020).

The first category (Table 3) includes those principles that should be considered for enabling managers to set up and then continuously manage the AI ecology: *design for shared norms* (DP1) and *design for human intervention* (DP2). Here, the focus is on the deliberate encoding of deep and surface norms and on monitoring outcomes and being able to intervene. Note our previous discussion of the interrelationships between the roles of managers and developers: Managers will often rely on developers to encode deep norms if this involves, for instance, training using data generated through simulation. In some situations, managers will have the required development capabilities.

The second category (Table 4) includes principles for designing agents so that they can effectively enact deep norms in known situations, develop new coordinated interactions and eventually norms in unknown situations, ensure consistency with established deep norms, and reinforce successful new norms: *design for predictability through sensory signals* (DP 3) and *design for predictability by enacting norms* (DP 4).

Table 3. Design Principles Focusing on Enabling Managerial Intervention in Norm-Based Coordination in Human-AI Ecologies

Principle	Subprinciple	Mechanism	Roles in implementation	Intended coordination outcome
DP1. Design for shared norms	**DP1.1 Deep norm encoding** Managers should be able to encode deep norms that reflect the ecology's values in coordination scenarios.	Values of a human-AI ecology should be implemented as shared high-level predictions with expected high precision so that possible violations lead to high prediction errors as agents receive sensory signals reflecting relevant contextual cues.	***Developers*** implement features for deep norm encoding, e.g., for training. ***Managers*** decide on deep norms. ***Developers*** work with *managers* to encode deep norms, e.g., through training. ***Agents'*** hierarchical prediction models are updated.	Human and AI-based agents draw on the same deep norms (prior knowledge). Both types of agents operate in the same environment (and hence encounter shared cues). Both types of agents evaluate if surface norms and coordinated interactions remain within desired boundaries.
	DP1.2 Selective surface norm encoding Managers should be able to encode desired surface norms that agents select by default when they encounter specific coordination scenarios.	Surface norms can be specified in terms of lower-level predictions with expected high precision about coordination scenarios, such that agents predict their interactions and incoming sensory signals reflecting relevant contextual cues in known situations.	***Developers*** implement features for surface norm encoding, e.g., for training. ***Managers*** decide on surface norms. ***Developers*** work with managers to encode surface norms, e.g., through training. ***Agents'*** hierarchical models are updated.	Human and AI-based agents enact shared surface norms when they recognize a coordination scenario with well-known features. As both types of agents enact the same surface norm, they engage in coordinated interaction to accomplish a desired outcome.
DP2. Design for human intervention	**DP2.1 Outcome monitoring and intervention** Managers should be able to monitor the emergence of surface norms and how their outcomes impact coordinated interactions with other agents.	Managers are provided with means to gain awareness of prediction errors as well as associated sensory signals and related contextual cues resulting from the violation of deep norms through the enactment of surface norms.	***Developers*** provide features for monitoring and intervention. ***Agents*** interact and create new behaviors, leading to new norms. ***Developers*** continuously monitor the emergence of new surface norms and identify violations of deep norms.	Managers intervene when surface norms violate deep norms, as observed by the coordinated interactions of human and AI-based agents.
	DP2.2 Deep norm reinforcement Managers should be able to reinforce surface norms as deep norms.	When surface norms reliably lead to desired interactions, managers can embed them as shared high-level deep norms such that future surface norms will emerge on the grounds of these norms.	***Developers*** implement features that allow managers to label successful norms. ***Managers*** work with *developers* to reinforce successful surface norms that help meet the ecology's goals, for instance, through training. ***Agents'*** hierarchical models are updated as successful surface norms receive increasingly high precision weights.	Desired new surface norms emerging from the interaction of human and AI-based agents become part of agents' action repertoires—and thus of the overall human-AI ecology—as they are assigned high precision.

Table 4. Design Principles Focusing on Enabling Agent Interaction in Norm-Based Coordination in Human-AI Ecologies

Principle	Subprinciple	Mechanism	Roles in implementation	Intended coordination outcome
DP3. Design for predictability through sensory signals	**DP3.1 Norm-enactment communication** An agent should be able to signal to other agents what surface norm it is enacting in a given coordination scenario.	Agents should convey sensory signals that serve as contextual cues with high precision weights for intended surface norm-based behavior for other agents.	***Managers*** work with *developers* to decide on relevant communication features. ***Developers*** provide *agents* with communication features for signaling relevant coordinated interactions. ***Agents***, then, communicate their coordination interactions to other agents in a given coordination scenario.	Human and AI-based agents both provide sensory signals for other agents to identify the norm they are following and, in turn, select appropriate behavior (i.e., make better predictions about the respective other agents' interactions in a coordination scenario).
	DP3.2 Norm-enactment inference An agent should be able to infer what surface norm another agent is enacting in a specific coordination scenario based on sensory signals.	Agents should recognize sensory signals by other agents as contextual cues to predict the intended interaction of the other agents with expected high precision.	***Managers*** work with *developers* to identify relevant contextual cues that reflect coordinated interactions related to relevant norms. ***Developers*** provide *agents* with sensory features for identifying cues in the environment and predicting other agents' coordinated interaction (i.e., identify an enacted norm) ***Agents***, then, receive sensory input and make predictions about other agents' interactions in a given coordination scenario.	Human and AI-based agents are able to make inferences about the surface norm associated with the behaviors of other agents to select actions that align with those of other agents while they simultaneously pursue their own goals.
DP.4 Design for predictability by enacting norms	**DP4.1 Surface norm defaulting (in known coordination scenarios)** When an agent faces a known coordination scenario, it should be able to enact an existing surface norm.	When an agent faces a coordination scenario, it should search for and give preference to the enactment of an existing prediction about a surface norm that has the highest expected precision in light of sensory signals that reflect relevant contextual cues.	***Developers*** provide *agents* with features to choose from competing predictions to minimize prediction error. ***Agents***, then, choose existing norms that are predicted to minimize prediction error in a coordination scenario based on sensory cues present in a given coordination scenario.	Human and AI-based agents enact the same surface norm—and thus enact the same coordinated interactions—when they encounter coordination scenarios where respective surface norms led to desired results in the past.

	DP4.2 Surface norm adjustment (in low-uncertainty situations) When an agent faces an unknown coordination scenario with low uncertainty, it should be able to adjust an existing surface norm.	When an agent encounters an unknown coordination scenario, it should enact and adjust a prediction about an existing surface norm with the highest expected precision when it does not violate a deep norm.	***Developers*** provide *agents* with features to adjust predictions about existing surface norms based on feedback they receive from the coordinated interaction in a given situation. ***Agents***, then, update their prediction models based on the feedback they receive.	Human and AI-based agents recognize coordination scenarios that they have not encountered before. In such situations, they agree on and adjust established surface norms to engage in coordinated interactions.
	DP4.3 Surface norm generation (in high-uncertainty situations) When an agent faces an unknown coordination scenario with high uncertainty, it should be able to generate a new surface norm.	When an agent encounters an unknown coordination scenario where no existing surface norm can be applied, it should generate a new coordinated interaction that can become a surface norm when the prediction error does not indicate a violation of a deep norm.	***Developers*** provide *agents* with features to generate new coordinated interactions and features for evaluating associated outcomes. ***Agents****, then,* make predictions about potential coordinated interactions in unknown situations and select those interactions for coordination that minimize prediction error.	Human and AI-based agents recognize coordination scenarios that are new and cannot be approached with established surface norms. On the grounds of their shared awareness of deep norms, they generate new coordinated interactions that, if recurrently enacted, become surface norms.
	DP4.4 Surface norm reinforcing (for future scenarios) When agents adjust or generate a surface norm in an unknown coordination scenario, they should reinforce it for future scenarios.	Surface norms adjusted or generated in light of uncertain coordination scenarios should be assigned high expected precision for future situations, as long as they do not produce prediction errors related to a deep norm.	***Developers*** provide *agents* with features to assign high precision weights to new coordinated interactions that do not violate deep norms. ***Agents****,* then, assign high precision to new coordinated interactions that do not violate deep norms.	Human and AI-based agents keep enacting generated surface norms in a given coordination scenario. Thereby, they reinforce these surface norms and engage in coordinated interactions.

DP1. Design for Shared Norms

Grounded in our analysis of the relation between deep norms, surface norms, and coordinated interactions, our design principles suggest that human-AI ecologies should provide a managerial interface that allows managers to influence the human-AI ecology with the help of developers by encoding shared norms to guide the coordinated interactions of agents.

DP1.1 Deep norm encoding: For human-AI ecologies to remain within desired boundaries, both human and AI-based agents need to draw on the same prior knowledge to generate shared predictions about desired and undesired outcomes of coordination scenarios. Deep norms ensure that the human-AI ecology is stable and evolves within "zones of desirable behavior" (Gasser & Mayer-Schönberger, 2024) when surface norms emerge. Deep norms can be based on societal values but also on laws and ethical standards (Gasser & Mayer-Schönberger, 2024). The key is that deep norms are based on a consensus about what forms of interactions should or should not occur between human and AI-based agents. In a warehouse ecology, for instance, a deep norm can require that AI-based agents must preserve human safety or handle items carefully to avoid damage (Blake, 2023).

In terms of predictive processing, deep norms are shared high-level predictions (Kiverstein & Rietveld, 2012; Williams, 2018). All encoded deep norms need to be assigned high precision, meaning that the enactment of a deep norm should be specified with regard to relevant contextual cues in a given coordination scenario (Clark, 2016). Hence, the general nature of the deep norm will be difficult to capture. This is a problem because the application of deep norms often requires causal extrapolation—which current generations of AI-based, predictive machines are incapable of (Schölkopf, 2022). For instance, reinforcement-based learning requires events to take place to train prediction, but this is often not possible. In the warehouse ecology example, we do not want to hurt humans in training the model to avoid these consequences.

However, managers and developers can train prediction models to avoid contacting humans and can also train models to recognize anomalous events. They can establish rules for proxies of adverse events and can further ensure the recognition of those adverse events through high prediction errors. One approach involves establishing a well-trained rule for action, and the other is a guide for dealing with unexpected outcomes. Thus, we propose drawing on general ethical systems as a guide for encoding deep norms. A key distinction recognized in the information systems literature is between deontological, or rule-based, standards of ethics and teleological, or consequentialist, standards (Berente et al., 2011). By combining these two standards, one can encode deep norms to flexibly deal with situations. To address known adverse outcomes, one can train models to deal with proxies for those outcomes in a way in which explicit rules are implicitly embedded in hierarchical prediction models. For example, one can train an agent on the rule to avoid human contact, where human contact is a proxy for harm. Second, for those unpredictable situations where proxies are not readily known, models can be trained to act on anomalous consequences. For example, if the location or position of humans in the ecology is anomalous—marked by a higher prediction error—then the deep norm can invoke surface norms to guide coordinated interactions, such as stopping, inquiring, and alerting.

In any case, the deep norm is either "hard-coded" or trained as a relationship between the agent's interactions and incoming sensory signals reflecting contextual cues that should be present or absent in a given situation. For instance, the deep norm that AI-based agents must not collide with each other or with humans (Schreiber, 2019) is reflected in the relationship between the agent's current interaction (e.g., movement) and incoming sensory signals that reflect relevant cues from the environment (e.g., parts of the human worker's body). Through training, agents learn upfront how, why, and when contextual cues reflect situations where humans are in danger. This is achieved, for example, when AI-based agents are trained in simulations where they are repeatedly exposed to situations where their interactions lead to desired and undesired outcomes. This involves the creation of synthetic data sets with undesired scenarios (e.g., a human is harmed) where the agent learns to establish correlations between antecedents, consequences, and influencing factors involved in undesired situations. Through such simulation-based training, predictions about deep norms are equipped with expected high precision, and a potential violation will be recognized as highly weighted prediction errors (Clark, 2016). When operating in a warehouse ecology, incoming sensory signals can cause the agent to pause or adjust its current interactions depending on hard-coded surface norms.

DP1.2 Selective surface norm encoding: Whereas deep norms induce stability in human-AI ecologies by constraining the kinds of surface norms that emerge and thus providing guardrails for the ecology, from a managerial point of view, the possibility of guiding more specific behaviors is desirable. The deep norm that humans should not be harmed provides a broad guardrail but does not readily translate into specific coordinated interactions. Surface norms based on such deep norms provide these more specific guardrails. To this end, managers may seek to encode surface norms. Consider how, in the warehouse ecology, for certain items that contain hazardous substances (Inam et al., 2018), the agent should not simply hand such items to another agent but should instead transfer them in a specific way to ensure safety. In such cases, the abstract deep norm alone is not sufficient because it could cause the emergence of surface norms that would entail handovers and interactions with such items in potentially dangerous ways. Hence, when both human and AI-based agents encounter a coordination scenario espousing well-known features, both types of agents should engage in coordinated interactions to ensure desired outcomes. Humans can also be educated and reminded about surface norms (Lupyan & Clark, 2015).

From the perspective of predictive processing, an encoded surface norm is implemented in AI-based agents as a prediction about the agent's specific coordinated interactions with expected high precision. Similar to implementing a deep norm, a surface norm is encoded in AI-based agents by training agents to recognize certain contextual cues and associate them with particular coordinated interactions. For instance, a deep norm specifying that the AI-based agent should not collide with humans in order to preserve their safety could be associated with an encoded surface norm specifying that the agent should stop, pause all interactions, and call for human supervision when contextual cues indicate that a human worker is injured. How surface norms are encoded in AI-based agents resembles the procedure through which deep norms are embedded in the agent: Through training, explicit plans are turned into implicit patterns in the agents' models. The key difference is the sequence in which the training takes place. First, deep norms are encoded; subsequently, surface norms are encoded under consideration of deep norms.

DP2. Design for Human Intervention

To ensure that agents preserve the values of the human-AI ecology, managers must be aware of how coordinated interactions emerge between agents (Metcalf et al., 2019) and be able to intervene when undesired or potentially dangerous surface norms emerge. Furthermore, managers should be enabled to adjust deep norms over time such that the stability of the human-AI ecology is retained.

DP2.1 Outcome monitoring and intervention: Over time, agents in human-AI ecologies produce surface norms based on recurrent behaviors in order to engage in coordinated interactions. While any surface norm should entail values embedded in deep norms, the outcomes of applying surface norms must be consistently and continuously monitored. It may be the case that agents accommodate violations of deep norms over time or that the environment in human-AI ecologies changes such that individual violations of deep norms are not recognized based on relevant contextual cues. Human agents might be aware of the predictions that appear correct or incorrect in a given situation, or they can be notified when their interactions deviate and appear problematic through some form of alert. However, AI-based agents may adopt undesired surface norms without reasoning about their appropriateness. Hence, managers should closely monitor the emergence of surface norms and the outcomes they produce, particularly by AI-based agents.

In terms of predictive processing, surface norms might lead to undesired outcomes in two ways. On the one hand, small violations of deep norms might lead to continuous, weak prediction errors. Over time, agents might accommodate and ignore these prediction errors, thus leading to the risk of undesired interactions being normalized (i.e., becoming part of a surface norm). For instance, while a deep norm may specify that an AI-based agent should not collide with humans, it might be the case that the agent gradually reduces the desired distance to humans over time, such as when operations are simultaneously made more efficient. Clearly, since the deep norm is represented as a high-level prediction with expected high precision, the agent will encounter prediction errors when the distance to humans decreases. Yet when the training underlying the deep norm is gradually overwritten, the agent may accommodate prediction errors over time. In such cases, managers must intervene.

On the other hand, the environment in a human-AI ecology can change such that certain contextual cues become relevant yet cannot be recognized by an AI-based agent because it has not been trained to do so. Consider, for example, how humans may informally start using ladders to retrieve items from higher up on the shelf. The deep norm of not colliding with humans has been trained under the assumption that agents can always recognize contextual cues that reflect parts of the human body, such as legs and feet. If AI-based agents do not associate ladders with the possibility of human workers standing on them, they might develop surface norms where their interactions are (too) close to ladders. As managers monitor the human-AI ecology, they can recognize these instances and adjust norms accordingly.

Therefore, it is important for AI-based agents in human-AI ecologies to not only be trained to project high precision on the grounds of sensory signals representing relevant contextual cues (Kiverstein & Rietveld, 2012); rather, they should also be designed to collect and process a large variety of sensory signals. With more sensory signals collected in a given situation and better AI-related training to map these sensory signals to desired and undesired outcomes, the expected precision of the agent's predictions can be increased. When an agent is equipped with the right sensors, it can potentially encounter a wider spectrum of prediction errors—indicating, for example, that a human worker is on a ladder and predicting the amount of danger the worker is in. For instance, the agent could consider the angle formed by the ladder in conjunction with the surface on which it stands (i.e., a wider angle between a ladder and a shelf may be more dangerous if that ladder is standing on tiles) or cues suggesting the physical condition of the worker (e.g., a worker's shaky movement might indicate that the worker is struggling, which could increase the likelihood of an accident). A crucial implication is that sensors should be embedded in AI-based agents that take the specific human-AI ecology into account—that is, what the environment looks like, what tasks are being performed, and what other agents are (likely) involved.

DP2.2 Deep norm reinforcement: As the human-AI ecology evolves over time, new surface norms will emerge between human and AI-based agents. As managers monitor the emergence of new surface norms, they may observe that some of them are particularly desirable. For instance, drones may join a warehouse ecology as a new class of AI-based agents to perform various tasks (Wawrla et al., 2019). In response, agents may develop surface norms to interact with drones, which might specify that an AI-based agent should not hand over any item when a human agent is within a radius of 5 meters. While this surface norm might emerge based on the deep norm of not harming human workers, managers might also realize the importance of implementing a generalized form of this norm whenever drones are present. Hence, the surface norm is reinforced and becomes a deep norm in AI-based agents that specifies no interaction with a flying drone when a human is proximal.

From a predictive processing point of view, surface norms are reinforced as deep norms when they are translated as abstract, generalizable, high-level predictions with expected high precision that are shared across human and AI-based agents in the human-AI ecology (Clark, 2016). Hence, contextual cues reflecting the features of a given coordination scenario will be assigned high precision weights (Kiverstein & Rietveld, 2012).

Regarding the design of AI-based agents, managers might reinforce the surface norm as a deep norm by considering all AI-based agents in the warehouse. As a result, the drones will likely adopt this norm because there is no other interaction possible with agents in the warehouse. Depending on their sphere of influence, managers might also embed this deep norm in other agents, e.g., drones. Furthermore, managers might design AI-based agents such that they continuously strengthen those deep norms that remain reliable over time and across coordination scenarios. From the viewpoint of predictive processing, agents reinforce predictions with expected high precision in relation to relevant contextual cues (Clark, 2016). For this, it is critical that the agent consider those sensory signals that may be reflective of whether the deep norm leads to desired coordination scenarios or provokes harmful situations.

DP3. Design for Promoting Predictability Through Sensory Signals

It is important that agents' intended interactions are predictable. Predictability implies that agents can select predictions about what is likely to happen next based on contextual cues transmitted by other agents. Hence, predictability is enabled in two ways. On the one hand, agents should be able to convey sensory signals about their intended interactions; on the other hand, they should be able to infer how other agents will interact based on the sensory signals transmitted by those agents.

DP3.1 Norm-enactment communication: To enhance predictability among human and AI-based agents, both types of agents need to communicate relevant signals indicative of planned coordinated interactions. This is particularly important for AI-based agents because they need to be explicitly designed to make themselves predictable to other agents by signaling the surface norm they intend to enact in a given coordination scenario. This helps other agents predict what they should do to enable coordinated interaction. In warehouse ecologies, for instance, an AI-based agent might signal the surface norm of handing over a heavy package by displaying sensory signals (e.g., flashing lights) (Johnson, 2022). When another agent recognizes these signals, it will enact the shared surface norm and align its interactions to receive an item that is potentially heavy. From the viewpoint of predictive processing, these signals are assigned high precision weights. These are either manually assigned to AI-based agents by the developer upfront, or they are repeatedly enacted such that the transmitted contextual cues become associated with incoming sensory signals that make the intended behavior of the agent reliably predictable to other agents (Williams, 2018). Furthermore, agents should enact surface norm-based interactions such that other agents can assign high precision weights. This requires that they take actions that other agents are able to detect, for instance, through their sensors. Furthermore, agents should be designed in a way that they enact interactions such that they remain consistent across coordination scenarios. When an agent interacts in a very similar way across coordination scenarios, other agents are more likely to associate this interaction—along with the transmitted sensory signals—with a surface norm.

DP3.2 Norm-enactment inference: Human and AI-based agents should be able to infer the intended surface norm-based interactions of other agents from the contextual signals transmitted by these agents. This enables them to choose interactions that allow them to coordinate with other agents. For instance, when an agent observes another agent displaying a red light as it slows down, it might predict that the other agent is enacting a surface norm signaling that it is approaching an ambiguous or even dangerous coordination scenario. In response, the agent will slow down its own operations and maintain distance between the other agent. Seen from the perspective of predictive processing, the agent can detect contextual cues as incoming sensory signals with high precision weights (Clark, 2016) that reliably reflect the intended surface norm-based behavior of other agents. As human agents are exposed to recurrent coordinated interactions, they learn to associate certain sensory signals reflecting relevant contextual cues with high precision weights (Linson et al., 2018). AI-based agents, however, need to be designed accordingly.

From the perspective of predictive processing, the AI-based agent is able to recognize these cues when it assigns high precision weights to these signals. This, in turn, can be provided through manual encoding or training, but the agent can also learn through repeated interactions that certain sensory signals can be used to reliably predict the intended interactions of another agent. However, actions alone do not encompass all potential surface norms. In principle, any interaction can belong to an arbitrary number of surface norms, and, vice versa, any surface norm can lead to an arbitrary number of interactions. Humans are likely to contextualize interactions within "policies" whereby a distinct set of actions is recognized and adequately interpreted (Linson et al., 2018).

It is crucial that AI-based agents are designed so that they are able to contextualize interactions such that they can map them to specific norms. This can happen in two ways. First, they can be designed to be able to integrate the larger context to relate interactions to surface norms, such as the state of the warehouse environment (e.g., how busy it is or the time of day; see Kiverstein & Rietveld, 2012). Again, more sensory signals will make it more likely that the agent will draw appropriate inferences. Second, agents can be designed to associate interactions with specific surface norms when they occur in certain sequences. When an agent encounters another agent that moves more slowly than usual and subsequently increases its distance to the shelf, one agent may predict that the other recognizes a dangerous situation (e.g., items may fall from the shelf).

DP4. Design for Predictability by Enacting Norms

To maintain stability in the human-AI ecology, agents should be designed to enact existing surface norms whenever possible, develop new norms only when necessary, and keep and reinforce those norms that are successful.

DP4.1 Surface norm defaulting (in known coordination scenarios): When human and AI-based agents engage in coordinated interactions, they should give preference to existing surface norms. If all involved agents recognize the situation as being linked to a given surface norm, they will likely enact the same surface norm and maintain stable and efficient coordinated interactions. For instance, agents in a warehouse ecology might routinely hand over items when they proceed from one area to the next (Wurman et al., 2008). Over time, human and AI-based agents will have developed the surface norm that several items can be handed over from one agent to the next when these items do not exceed a certain size and can be handled safely (an underlying deep norm might be that items should be handled carefully). Whenever agents encounter a situation where they hand over one or more items, they enact the surface norm to hand over several items when the associated conditions are met. Human agents default to known coordinated interactions after they have learned them based on observation and repetition (Köster et al., 2020).

From the perspective of predictive processing, an existing surface norm should be selected by AI-based agents as a prediction with expected high precision among multiple competing predictions representing alternative surface norms. Repetition of a surface norm across situations decreases uncertainty. Expected high precision of a prediction, in turn, is linked to contextual cues of a given situation that are transmitted as incoming sensory signals (Clark, 2016). For instance, an agent will recognize that another agent moves in a way to accept more than one item in a hand-over situation, thus giving preference to the prediction of a surface norm that it can hand over multiple items.

DP4.2 Surface norm adjustment (in low-uncertainty situations): When a human or AI-based agent encounters a coordination scenario that has a low degree of uncertainty, it should adjust existing surface norms as long as they do not violate existing deep norms. For instance, an AI-based agent might seek to hand over a bulky item to a human worker. The item is new in the warehouse, and the agent has not encountered this situation before. The agent thus draws on an existing surface norm that has been enacted in previous situations where the agent signals to the human worker that it intends a handover. However, instead of directly handing over the item to the human worker and putting the human worker at risk (Silverman, 2019), it will adjust the existing surface norm and first put the item on the ground, thereby enacting an action that deviates from the existing norm. The human worker can now take over the item safely. In this case, the surface norm is adjusted in situ while the deep norm is preserved. Human agents are flexible and can adjust their behavior in light of changing environments (Clark, 2016). However, AI-based agents need to be designed accordingly.

From the viewpoint of predictive processing, the AI-based agent encounters a situation where contextual cues are transmitted as sensory signals that do not evoke a prediction about a surface norm with high precision. The situation yields uncertainty, and the agent is exposed to prediction errors (Clark, 2016). However, the prediction errors do not violate deep norms where specific contextual cues about the state of the human worker are assigned high precision weights. For instance, the AI-based agent is exposed to high prediction error upon approaching the human worker directly with the bulky item because it would put the human worker at risk. Hence, by avoiding prediction errors that potentially violate predictions about deep norms, the agent adjusts its surface norm.

DP4.3 Surface norm generation (in high-uncertainty situations): When human and AI-based agents encounter unknown coordination scenarios where they cannot adjust an existing surface norm, they should be able to generate a new surface norm considering existing deep norms. For instance, two agents might encounter each other in the hallway of the warehouse ecology. Usually, there are two lanes available for agents traveling in opposite directions. However, since new items have arrived in the warehouse, which could not yet all be stored on the shelves, there is only one available lane. The agents develop new behavior that, if reenacted over time, becomes a surface norm (see DP4.5 on reinforcement) whereby one agent continues to move along its lane (which is available) while the other agent stops and pauses until the other agent has passed. Again, human agents may adjust their predictions to improvise in novel situations (Clark, 2018).

In terms of predictive processing, AI-based agents should be designed to generate new surface norms when encountering unknown coordination scenarios with high uncertainty (Clark, 2016). In such situations, they cannot apply an existing surface norm in light of incoming sensory signals that reflect established contextual cues. As they generate a new surface norm, they develop an interaction pattern that can be predicted by both agents. Importantly, the generation of new surface norms hinges on shared deep norms, which serve two central functions. On the one hand, deep norms ensure that both agents agree on coordinated interactions that are in line with the values of the human-AI ecology. On the other hand, deep norms narrow down the possible range of surface norms. As all agents involved in a high-uncertainty situation share the same deep norms, they will recognize, for instance, the constraints of a given situation in terms of what can or must not be done.

DP4.4 Surface norm reinforcing (for future coordination scenarios): When human and AI-based agents adjust existing surface norms in low-uncertainty coordination scenarios or generate new ones in high-uncertainty ones, they should reinforce them in and for future coordination scenarios when they are successful and do not violate deep norms. For instance, as illustrated in DP4.2, agents should reinforce the surface norm that bulky items should be handed over by placing them on the ground first; the agent may even adjust surface norms in coordinated interactions with other agents by always placing bulky items on the ground. Similarly, as prescribed in DP4.3, agents should reinforce the surface norm that agents take turns moving forward when one lane in the warehouse hallway is blocked. Human agents tend to learn through successful interactions and thus strengthen surface norms over time (Köster et al., 2020). AI-based agents, however, need to be designed accordingly.

From the perspective of predictive processing, the reinforcement of surface norms occurs when AI-based agents repeatedly encounter the same situation. As they encounter contextual cues representing a known coordination scenario, predictions about surface norms are enacted with higher expected precision. Again, repetition and reinforcement are key (Clark, 2016), especially when training AI-based agents. Reenacted surface norms travel across the ecology and become widely available to agents. When agents recurrently face similar situations where the same surface norms lead to desired effects, they will be more likely to reinforce these surface norms.

Discussion

In this paper, we have explored how human and AI-based agents coordinate in human-AI ecologies, combining predictive processing theory and social norms theory. Norms are a critical way to enact guardrails in these ecologies. They provide "zones of desirable behavior" in which agents interact over time (Gasser & Mayer-Schönberger, 2024), but norm-based coordination is challenging because agents face uncertainty and unstructured coordination scenarios where they cannot readily predict what they should do next (Claggett & Karahanna, 2018; Faraj & Xiao, 2006). In such cases, no norms might be available that directly apply to coordination needs; therefore, norms for coordination must be adapted or created. It is critical, then, that human values guide the establishment of norms (Claggett & Karahanna, 2018; Kane et al., 2021; Kissinger et al., 2021). Whereas values are reflected in humans through socialization and discretion, AI-based agents do not experience socialization in human communities; therefore, they do not have the associated discretion to ensure that emerging norms are consistent with appropriate values (Bullock, 2019). Against this background, our paper proposes design principles to guide the development of norms for coordination in human-AI ecologies, which adhere to values and thus prevent the emergence of undesired outcomes. The emphasis of this work involves design for predictability. Norms enable coordination among diverse agents in no small part because they render behavior predictable. It is critical as we design and manage human-AI ecologies that we embrace this predictability.

Implications for AI Policymaking and Regulation

Managing norm-based coordination in human-AI ecologies is implied in AI policymaking and regulation in two main ways: top-down, as organizations seek to implement policies in their design and management of human-AI ecologies, and bottom-up, as policymakers and regulators identify a mismatch between emergent AI-enabled behaviors and norms and current regulatory design.

First, our design principles can support organizations implementing policies and regulatory requirements by providing explicit mechanisms for developers and managers to embed regulatory requirements in their design of AI ecologies, including the design of specific agents. By embedding deep and surface norms in hierarchical prediction models, AI-based agents in human-AI ecologies can serve as a means for implementing regulatory principles and rules. The differentiation between deep norm encoding and surface norm adjustment, generation, and reinforcement, for instance, is reflective of the idea that regulation in the face of sociotechnical change (Bennett Moses, 2016) should be principle-based[4] (e.g., Carter & Marchant, 2011) and allow for a variety of implementations. General principles that must not be violated can be embedded as deep norms in human-AI ecologies, but they can then materialize in various behaviors reflected in surface norms compliant with those deep norms. Consider how the EU AI Act provides basic principles reflecting core values: human agency and oversight; technical robustness and safety; privacy and data governance, transparency, non-discrimination, and fairness; social and environmental well-being; and accountability (European Parliament and Council, 2024). These are examples of values that developers and managers will have to collaborate on to embed in human-AI ecologies as deep norms. Further, equipping managers with means to continuously manage AI ecologies can help implement post-market monitoring systems, such as required in Article 72 of the EU AI Act, which states that "providers shall establish and document a post-market monitoring system in a manner that is proportionate to the nature of the AI technologies and the risks of the high-risk AI system" and that "the post-market monitoring system shall actively and systematically collect, document and analyze relevant data which may be provided by deployers or which may be collected through other sources on the performance of high-risk AI."

Regulation in the face of sociotechnical change (Bennett Moses, 2016) seeks to avoid harmful effects while being innovation-friendly (Butler et al., 2023). Achieving these goals requires policymakers to make sense of emerging technologies and their uses (Seidel et al., 2025). To this end, policymakers and regulators can use the monitoring of emergent norms in AI ecologies to identify new uses of technology, monitor the suitability of current regulations (Baldwin et al., 2011), and identify regulatory gaps (Mandel, 2009). This information can inform the sensemaking process that is required in regulating emerging technologies, as regulators, for instance, seek to abstract from specific behaviors to be able to regulate technologies and their uses in an implementation-neutral way (Seidel et al., 2025), that is, in a way that does not discriminate against any specific technology (Koops, 2006). Thus, norms that are enacted in AI ecologies are proxies for the behaviors enabled by AI-based technologies, with deep norms describing higher levels of abstraction than surface norms. In particular, emergent deep norms consistent with established deep norms can guide policymaking and regulatory design. In other words, norm emergence provides a logical point of departure for regulatory change.

[4] A general distinction is between principles-based and rule-based (Burgemeestre et al., 2009), and both principles and rules that must not be violated can be translated into deep norms.

Ultimately, we can conceive of a situation where the bottom-up and top-down aspects of regulation are increasingly dissolving. The relationship between those aspects that manage human-AI ecologies and those that manage policymakers are becoming blurred, as these ecologies permeate organizational boundaries, as seen in the case of autonomous transportation. Policymakers and regulators work with organizational managers to provide fundamental principles and rules that can be used in human-AI ecologies in the design phase as well as for regular updates. Recent EU regulations, such as the Markets in Crypto-Assets Regulation (European Commission, 2022; European Parliament and Council, 2022), require that specific information regarding the implementation of rules is available in machine-readable form. Regulatory agencies can then use monitoring outcomes—prediction errors and associated environmental cues—to revise policy on an ongoing basis. Clearly, this provides additional design challenges.

Implications for Information Systems Research

The prevailing assumption in information systems research is that humans and AI should be treated as two distinct kinds of agents. There is a division of labor assumed in the general responsibilities of these two types of agents and in the particular features and skills: AI-based agents excel at some types of tasks and human agents at others (Raisch & Fomina, 2025). A prevailing assumption is that it is the human who delegates tasks to AI (Baird & Maruping, 2021) and should be kept "in the loop" to correct AI-based decisions and actions (Fügener et al., 2021). Other research has highlighted that, while humans and machines engage in mutual model adaptation, their respective models are conceived as fundamentally distinct and in constant need of reconciliation (Seidel et al., 2018; Te'eni et al., 2023).

The argument we propose here complements—and to a certain extent challenges—these views by suggesting a more *symmetrical* view of humans and AI. Drawing on predictive processing theory, we assume that human and AI-based agents function in the same ways, at least in reference to basic information processing mechanisms (Clark, 2016). Both types of agents can be conceived of as Bayesian prediction machines that use previous experiences to predict what is likely to happen next, and they change and adjust their predictions when they encounter prediction errors. Certainly, this is a rather broad-brush approach that abstracts from important differences—for instance, that human cognition appears to be much more embodied (Clark, 2016) and that a variety of machine learning techniques exist (Russell & Norvig, 2020). In any case, the symmetry that this view proposes offers potential for theoretical development in two main areas.

First and foremost, the implied symmetry allows us to embrace the fact that AI-based agents have become increasingly sophisticated, have higher autonomy, and have taken over a wider range of tasks. These technologies are no longer mere recipients of human requests, but they participate in and even influence the social fabric that has long been thought to be exclusively reserved for humans. Their roles have become increasingly equivalent to those of humans. We see indications for this in all kinds of social contexts. For instance, there are many organizational settings where AI-based algorithms delegate to humans (Möhlmann et al., 2021; Stelmaszak et al., 2024). In other situations, AI-based agents less directly influence human action. Robots are increasingly taking over all kinds of tasks and roles in manufacturing and warehouse settings, and humans are being put under pressure to keep up with their work pace, often leading to potential physical and psychological harm (Blake, 2023). As children increasingly interact with voice-based assistants in their homes, their interaction patterns adjust and change, and their social etiquette can become less polite, even rude (Christakis, 2019). Language communities modify their communication patterns in response to language preferences espoused in large language models (Yakura et al., 2024), and human-to-human interaction changes when AI-based agents are introduced to communities (Traeger et al., 2020; Safadi et al., 2024). Information systems research is increasingly concerned with situations where both human *and* AI-based agents learn and mutually adjust each other's behaviors (Lyytinen et al., 2021; Seidel et al., 2018). In this context, predictive processing theory provides an opportunity to theorize about individual and social processes.

Second, emphasizing a symmetry between human and AI-based agents enables novel means for interventions. When humans and AI-based agents are considered to operate under the same information processing principles, it is possible to conceptualize overarching management interventions that equally consider both types of agents. Consider, for instance, the creative production of knowledge in organizations, a context where humans and AI-based agents often work together (Raisch & Fomina, 2025; Seidel et al., 2018). Recent research has shown how the (organizational) environment can have an enabling effect on knowledge production (Knight et al., 2025), and a symmetry approach can guide manipulations that simultaneously affect both types of agents. For instance, the environment in the organization can be designed such that it triggers prediction errors that will cause both types of agents to readjust their established models. Such considerations seem particularly relevant for impact-oriented research approaches, and a theory, such as predictive processing theory, seems particularly apt for design science research. Our view that emphasizes design for predictability draws upon predictive processing and provides an original perspective that can inform this design research.

Furthermore, our paper aligns with recent arguments suggesting that information systems researchers should be turning to ethical questions around the use of AI. It has been warned, for example, that human society may be marching into undesired futures where AI-based agents disempower human agents and take actions that go against human interests (Kane et al., 2021; Young et al., 2021). In consequence, there have been calls urging users of AI-based systems to become more aware of their implications and counteract potentially malicious tendencies (Kane et al., 2021; Young et al., 2021). Our perspective is a departure from the conventional view of interaction between human and AI-based agents. Instead, we emphasize how it is impossible to foresee all future interactions among diverse agents in human-AI ecologies. Even if managers and developers may know precisely what counts as desirable or undesirable behavior, one cannot hard-code those behaviors because, assuming ecologies allow for some flexibility, these behaviors will change. Our norm-based approach offers a middle ground between determinism and arbitrariness. Accordingly, values, ethics, and moral ideas can be implemented as stable components in a human-AI ecology; simultaneously, flexibility can be enabled in dynamically changing environments. Hence, newly emergent behavior will provide ideas for evolving deep norms. Through our emphasis on design for prediction, our design theory suggests when, how, and why managers and developers should interfere and take corrective actions. Similarly, our research adds to recent arguments that the role of ethics in AI should be understood as dynamically evolving and changing (Berente et al., 2024). As ethical decisions are being implemented, they can be tested and simulated by developing them through deep norms "from the bottom up." For instance, implementing values into deep norms can be the basis for simulations to predict when, how, and why certain surface norms will emerge.

Limitations and Conclusion

It is important to note that our design principles are future-oriented and intended to inform inquiries about the design of human-AI ecologies. By drawing on predictive processing as a unifying theory for both human and machine agents, we sought a level of generality that would allow us to communicate our basic ideas to some "other mind"

(Churchman, 1971)—the recipients of these principles, in terms of developers and managers of AI ecologies, but also the agents that will implement these ideas. By using predictive processing to assume increasing symmetry between human and AI-based agents, we revisit the fundamental idea that machines can implement faculties, such as inquiry (Churchman, 1971), in the context of human-AI ecologies.

There are at least two main design problems that immediately follow from our design theorizing and present important design challenges in themselves. The first is the *implementation problem*. While predictive processing allows for a unified perspective on human and machine learning and sensemaking, current machines are not yet capable of fully implementing predictive processing (Ciria et al., 2021). While we have provided some pointers on how, specifically, the abstracted mechanisms can be implemented, for instance, through reinforcement learning, the suggested mechanisms are not limited to specific technologies. The second is the *update problem,* which is only partially addressed in our theory. When we suggest that managers encode both deep and surface norms, this requires both technical means to change the inner workings of AI-based agents and means to communicate new norms to human agents. Retraining large numbers of agents will be challenging. However, the emergence of norms is addressed by ongoing interactions as agents learn from prediction errors in known and unknown coordination scenarios and reinforce successful norms. Norms are social structures that travel across agents, as these create collective shared mental models in our theory, captured by shared hierarchical prediction models.

Our theoretical development reflects both our sensemaking informed by norm theory and predictive processing about how prediction errors provide a key explanatory mechanism for how various types of agents can coordinate in human-AI ecologies. It is a step in thinking through the broader issue of ensuring guardrails for powerful AI-based systems in a variety of contexts (Gasser & Mayer-Schönberger, 2024). Our design principles can inform design choices that will generate prediction errors and lead to updates of our understanding of how human-AI ecologies should be designed to meet their goals while adhering to important values.

Acknowledgements

We thank our senior editor, Ann Majchrzak, for her invaluable guidance throughout the revision process. We further thank the associate editor and the three reviewers for their excellent comments and helpful suggestions. We are also indebted to colleagues who have provided important feedback on earlier versions of our paper: Ahmed Abbasi, Waldemar Kremser, Oliver Lukitsch, Christian Mahringer, Wojtek Przepiorka, and Anuschka Schmitt.

References

Abbink, K., Gangadharan, L., Handfield, T., & Thrasher, J. (2017). Peer punishment promotes enforcement of bad social norms. *Nature Communications, 8*(1), Article 609. https://doi.org/10.1038/s41467-017-00731-0

Abrutyn, S., & Carter, M. J. (2015). The decline in shared collective conscience as found in the shifting norms and values of etiquette manuals. *Journal for the Theory of Social Behaviour, 45*(3), 352-376. https://doi.org/10.1111/jtsb.12071

Argote, L. (1982). Input uncertainty and organizational coordination in hospital emergency units. *Administrative Science Quarterly*, 27(3), 420-434. https://doi.org/10.2307/2392320

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*(1), 315-341. https://doi.org/10.25300/MISQ/2021/15882

Baldwin, R., Cave, M., & Lodge, M. (2011). *Understanding regulation: theory, strategy, and practice*. Oxford University Press.

Bennett Moses, L. (2016). Regulating in the face of sociotechnical change. In R. Brownsword, E. Scotford, & K. Yeung (Eds.), *The Oxford handbook of law, regulation and technology* (pp. 573-596). Oxford University Press https://doi.org/10.1093/oxfordhb/9780199680832.013.49

Berente, N., Gal, U., & Hansen, S. (2011). Ethical implications of social stratification in information systems research. *Information Systems Journal, 21*(4), 357-382. https://doi.org/10.1111/j.1365-2575.2010.00353.x

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing articial intelligence. *MIS Quarterly, 45*(3), 1433-1450. https://doi.org/10.25300/MISQ/2021/16274

Berente, N., Kormylo, C., & Rosenkranz, C. (2024). Test-driven ethics for machine learning. *Communications of the ACM, 67*(5), 45-47. https://doi.org/10.1145/3633455

Blake, R. (2023). Rise of warehouse robots spurs efficiencies—and safety concerns. *Forbes*. https://www.forbes.com/sites/richblake1/2023/11/24/rise-of-warehouse-robots-spurs-efficiencies-and-safety-concerns/?sh=1663864740a3

Browne, J. T., & Garnham, I. (2022). Extended, distributed, and predictive: Sketches of a generative theory of interaction for HCI. In *Proceedings of the International Conference on Human-Computer Interaction*, pp. 17-24. https://doi.org/10.1007/978-3-031-19679-9_3

Bullock, J. B. (2019). Artificial intelligence, discretion, and bureaucracy. *The American Review of Public Administration, 49*(7), 751-761. https://doi.org/10.1177/0275074019856123

Burgemeestre, B., Hulstijn, J., & Tan, Y.-H. (2009). Rule-based versus principle-based regulatory compliance. In E. Francesconi, G. Borges, G. and C. Sorge, C. (Eds.), *Legal Knowledge and Information Systems* (pp. 37-46). IOS Press. https://doi.org/10.3233/978-1-60750-082-7-37

Butler, T., Gozman, D., & Lyytinen, K. (2023). The regulation of and through information technology: Towards a conceptual ontology for IS research. *Journal of Information Technology, 38*(2), 86-107. https://doi.org/10.1177/02683962231181147

Carroll, J. M. (1997). Human-computer interaction: psychology as a science of design. *Annual Review of Psychology, 48*(1), 61-83. https://doi.org/10.1006/ijhc.1996.0101

Carter, R. B., & Marchant, G. E. (2011). Principles-based regulation and emerging technology. In G. E. Marchent, B R. Allenby and J.

R. Herkert (Eds.), *The growing gap between emerging technologies and legal-ethical oversight: The pacing problem* (pp. 157-166). Springer. https://doi.org/10.1007/978-94-007-1356-7_10

Chopra, A., van der Torre, L., Verhagen, H., & Villata, S. (2018). *Handook of normative multiagent systems*. College Publishing.

Christakis, N. (2019). How AI will rewire us. *The Atlantic*. https://www.theatlantic.com/magazine/archive/2019/04/robots-human-relationships/583204/

Chung, A. C. A., & Rimal, R. N. R. R. N. (2016). Social norms: A review. *Review of Communication Research, 4*, 1-28. https://doi.org/10.12840/issn.2255-4165.2016.04.01.008

Churchman, C. W. (1971). *The design of inquiring systems: Basic concepts of systems and organization*. Basic Books.

Ciria, A., Schillaci, G., Pezzulo, G., Hafner, V. V., & Lara, B. (2021). Predictive processing in cognitive robotics: a review. *Neural Computation, 33*(5), 1402-1432. https://doi.org/10.1162/neco_a_01383

Claggett, J. L., & Karahanna, E. (2018). Unpacking the structure of coordination mechanisms and the role of relational coordination in an era of digitally mediated work processes. *Academy of Management Review, 43*(4), 704-722. https://doi.org/10.5465/amr.2016.0325

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181-204. https://doi.org/10.1017/S0140525X12000477

Clark, A. (2016). *Surfing uncertainty. Prediction, action and the embodied mind*. Oxford University Press.

Clark, A. (2018). A nice surprise? Predictive processing and the active pursuit of novelty. *Phenomenology and the Cognitive Sciences, 17*(3), 521-534. https://doi.org/10.1007/s11097-017-9525-z

Crowston, K. (1997). A coordination theory approach to organizational process design. *Organization Science, 8*(2), 157-175. https://doi.org/10.1287/orsc.8.2.157

de Freitas, N. C. A., & dos Santos Silva, M. P. (2017). Multiagent system for image mining. In *Multi-agent Systems*. IntechOpen. https://doi.org/10.5772/intechopen.69353

Den Ouden, H. E., Kok, P., & De Lange, F. P. (2012). How prediction errors shape perception, attention, and motivation. *Frontiers in Psychology, 3*, Article 548. https://doi.org/10.3389/fpsyg.2012.00548

Diekmann, A. (2022). Emergence of and compliance with new social norms: The example of the COVID crisis in Germany. *Rationality and Society, 34*(2), 129-154. https://doi.org/10.1177/10434631221092

European Commission. (2022). *Impact assessment, evaluation and monitoring of EU research and innovation programmes*. https://research-and-innovation.ec.europa. eu/strategy/support-policy-making/shaping-eu-research-and-innovation-policy/evaluation-impact-assessment-and-monitoring_en?2nd-language=de

European Parliament and Council. (2022). *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services (Digital Services Act) (Vol. L277)*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065

European Parliament and Council. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (Vol. L279)*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689

Evans, W. (2020). *How Amazon hid its safety crisis*. Reveal. https://revealnews.org/article/how-amazon-hid-its-safety-crisis/

Faraj, S., & Xiao, Y. (2006). Coordination in fast-response organizations. *Management Science, 52*(8), 1155-1169. https://doi.org/10.1287/mnsc.1060.0526

Feldman, D. C. (1984). The development and enforcement of group norms. *Academy of Management Review, 9*(1), 47-53. https://doi.org/10.2307/258231

Frantz, C., & Pigozzi, G. (2018). Modeling norm dynamics in multiagent systems. In A. Chopra, L. van der Torre, H. Verhagen and S. Villata (Eds.), *Handbook of Normative Multiagent Systems* (pp. 491-564). College Publishing.

Frauenberger, C. (2019). Entanglement HCI the next wave? *ACM Transactions on Computer-Human Interaction, 27*(1), 1-27. https://doi.org/10.1145/336499

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: a process theory. *Neural Computation, 29*(1), 1-49. https://doi.org/10.1162/NECO_a_00912

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. *MIS Quarterly 45*(3), 1527-1556. https://doi.org/10.25300/MISQ/2021/16553

Gasser, U., & Mayer-Schönberger, V. (2024). *Guardrails: Guiding human decisions in the age of AI*. Princeton University Press.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Lawrence Erlbaum Associates.

Gittell, J. (2002). Coordinating mechanisms in care provider groups: Relational coordination as a mediator and input uncertainty as a moderator of performance effects. *Management Science, 48*(11), 1408-1426. https://doi.org/10.1287/mnsc.48.11.1408.268

Gregor, S., Chandra Kruse, L., & Seidel, S. (2020). Research perspectives: the anatomy of a design principle. *Journal of the Association for Information Systems, 21*(6), 1622-1652. https://doi.org/10.17705/1jais.00649

Harrison, S., Sengers, P., & Tatar, D. (2011). Making epistemological trouble: Third-paradigm HCI as successor science. *Interacting with Computers, 23*(5), 385-392. https://doi.org/10.1016/j.intcom.2011.03.005

Harrison, S., Tatar, D., & Sengers, P. (2007). The three paradigms of HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Hawkins, R. X., Goodman, N. D., & Goldstone, R. L. (2019). The emergence of social norms and conventions. *Trends in Cognitive Sciences, 23*(2), 158-169. https://doi.org/10.1016/j.tics.2018.11.003

Hohwy, J. (2013). *The predictive mind*. Oxford University Press.

Horne, C., & Mollborn, S. (2020). Norms: An integrated framework. *Annual Review of Sociology, 46*, 467-487. https://doi.org/10.1146/annurev-soc-121919-054658

Iacono, S. L., Przepiorka, W., Buskens, V., Corten, R., & van de Rijt, A. (2021). COVID-19 vulnerability and perceived norm violations predict loss of social trust: A pre-post study. *Social Science & Medicine, 291*, Article 114513. https://doi.org/10.1016/j.socscimed.2021.114513

Inam, R., Raizer, K., Hata, A., Souza, R., Forsman, E., Cao, E., & Wang, S. (2018). Risk assessment for human-robot collaboration in an automated warehouse scenario. In *Proceedings of the 23rd*

International Conference on Emerging Technologies and Factory Automation. https://doi.org/10.1109/ETFA.2018.8502466

Jennings, N. R., Moreau, L., Nicholson, D., Ramchurn, S., Roberts, S., Rodden, T., & Rogers, A. (2014). Human-agent collectives. *Communications of the ACM, 57*(12), 80-88. https://doi.org/10.1145/2629559

Johnson, K. (2022). Amazon's "safe" new robot won't fix its worker injury problem. *Wired*. https://www.wired.com/story/amazons-worker-injury-problem/

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*(1), 371-396. https://doi.org/10.25300/MISQ/2021/1578

Kelly, M. P., Kriznik, N. M., Kinmonth, A. L., & Fletcher, P. C. (2019). The brain, self and society: a social-neuroscience model of predictive processing. *Social Neuroscience, 14*(3), 266-276. https://doi.org/10.1080/17470919.2018.1471003

Kissinger, H., Schmidt, E., & Huttenlocher, D. (2021). *The age of AI and our human future*. Little, Brown & Company.

Kiverstein, J., & Rietveld, E. (2012). Dealing with context through action-oriented predictive processing. *Frontiers in Psychology, 3*, Article 421. https://doi.org/10.3389/fpsyg.2012.00421

Knight, E., Lok, J., Jarzabkowski, P., & Wenzel, M. (2025). Sensing the room: the role of atmosphere in collective sensemaking. *Academy of Management Journal*., *68*, 326-356 https://doi.org/10.5465/amj.2021.1389

Koops, B.-J. (2006). Should ICT regulation be technology-neutral? In B. Koops, M. Lips, C. Prins and M. Schellekens (Eds.), *Starting points for ICT regulation. Deconstructing prevalent policy one-liners, it & law series* (p. 77-108). Springer.

Köster, M., Kayhan, E., Langeloh, M., & Hoehl, S. (2020). Making sense of the world: infant learning from a predictive processing perspective. *Perspectives on Psychological Science, 15*(3), 562-571. https://doi.org/10.1177/1745691619895071

Kuutti, K., & Bannon, L. J. (2014). The turn to practice in HCI: towards a research agenda. In *Proceedings of the SIGCHI conference on human factors in computing systems*. https://doi.org/10.1145/2556288.2557111

Lanillos, P., & Cheng, G. (2018). Adaptive robot body learning and estimation through predictive coding. In *Proceedings of the International Conference on Intelligent Robots and Systems*. https://doi.org/10.1109/IROS.2018.8593684

Lewis, D. (1969). *Convention*. Harvard University Press.

Limerick, H., Coyle, D., & Moore, J. W. (2014). The experience of agency in human-computer interactions: A review. *Frontiers in human neuroscience, 8*, Article 643. https://doi.org/10.3389/fnhum.2014.00643

Linson, A., Clark, A., Ramamoorthy, S., & Friston, K. (2018). The active inference approach to ecological perception: General information dynamics for natural and artificial embodied cognition. *Frontiers in Robotics and AI, 5*, Article 21. https://doi.org/10.3389/frobt.2018.00021

Lobo, L., Heras-Escribano, M., & Travieso, D. (2018). The history and philosophy of ecological psychology. *Frontiers in Psychology, 9*, Article 2228. https://doi.org/10.3389/fpsyg.2018.02228

Loomba, S., De Figueiredo, A., Piatek, S. J., De Graaf, K., & Larson, H. J. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour, 5*(3), 337-348. https://doi.org/10.1038/s41562-021-01056-1

Lupyan, G., & Clark, A. (2015). Words and the world: Predictive coding and the language-perception-cognition interface. *Current Directions in Psychological Science, 24*(4), 279-284. https://doi.org/10.1177/0963721415570732

Lyytinen, K., Nickerson, J. V., & King, J. L. (2021). Metahuman systems=humans+machines that learn. *Journal of Information Technology, 36*(4), 427-445. https://doi.org/10.1177/02683962209159

Mandel, G. N. (2009). Regulating emerging technologies. *Law, Innovation and Technology, 1*(1), 75-92.

Marmor, A. (2009). *Social conventions—From language to law*. Princeton University Press.

Metcalf, L., Askay, D. A., & Rosenberg, L. B. (2019). Keeping humans in the loop: Pooling knowledge through artificial swarm intelligence to improve business decision making. *California Management Review, 61*(4), 84-109. https://doi.org/10.1177/0008125619862256

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(4)*, 1999-2022. https://doi.org/10.25300/MISQ/2021/15333

Morales, J., Lopez-Sanchez, M., Rodriguez-Aguilar, J. A., Wooldridge, M. J., & Vasconcelos, W. W. (2013). *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems* (483-490).

Morris, B. (2024). *How robots are taking over warehouse work.* BBC. https://www.bbc.com/news/business-68639533

Okhuysen, G. A., & Bechky, B. A. (2009). Coordination in organizations: An integrative perspective. *Academy of Management Annals, 3*(1), 463-502. https://doi.org/10.1080/19416520903047533

Przepiorka, W., Szekely, A., Andrighetto, G., Diekmann, A., & Tummolini, L. (2022). How norms emerge from conventions (and change). *Socius, 8*, 1-16. https://doi.org/10.1177/23780231221124556

Quinlivan, J. (2023). *How Amazon deploys collaborative robots in its operations to benefit employees and customers*. Amazon. https://www.aboutamazon.com/news/operations/how-amazon-deploys-robots-in-its-operations-facilities

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., & Jackson, M. O. (2019). Machine behaviour. *Nature, 568*, 477-486. https://doi.org/10.1038/s41586-019-1138-y

Raisch, S., & Fomina, K. (2025). Combining human and artificial intelligence: Hybrid problem-solving in organizations. *Academy of Management Review, 50,* 441-464. https://doi.org/10.5465/amr.2021.0421

Reichardt, R., Polner, B., & Simor, P. (2020). Novelty manipulations, memory performance, and predictive coding: The role of unexpectedness. *Frontiers in Human Neuroscience, 14*, Article 152. https://doi.org/10.3389/fnhum.2020.00152

Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th Edition). Prentice Hall.

Safadi, H., Lalor, J. P., & Berente, N. (2024). The effect of bots on human interaction in online communities. *MIS Quarterly, 48*(3), 1279-1296. https://doi.org/10.25300/MISQ/2023/17901

Savarimuthu, B. T. R., & Cranefield, S. (2011). Norm creation, spreading and emergence: A survey of simulation models of norms in multi-agent systems. *Multiagent and Grid Systems, 7*(1), 21-54. https://doi.org/10.5555/3635637.3663046

Schölkopf, B. (2022). Causality for machine learning. In H. Geffner, R. Dechter and J. Y. Halpern (eds.) *Probabilistic and causal inference: The works of Judea Pearl* (pp. 765-804). ACM Books. https://doi.org/10.1145/3501714

Schreiber, N. (2019). Inside an Amazon warehouse, robots' ways rub off on humans. *New York Times*. https://www.nytimes.com/2019/07/03/business/amazon-warehouse-labor-robots.html

Seidel, S., Berente, N., Lindberg, A., Lyytinen, K., & Nickerson, J. V. (2018). Autonomous tools and design: A triple-loop approach to human-machine learning. *Communications of the ACM*, 62(1), 50-57. https://doi.org/10.1145/3210753

Seidel, S., Frick, C. J., & vom Brocke, J. (2025). Regulating emerging technologies: Prospective sensemaking through abstraction and elaboration. *MIS Quarterly, 49*(1), 179-204. https://doi.org/10.25300/MISQ/2024/18039

Shaheen, S. (2023). Will the dark warehouse ever become reality? Perhaps not in our lifetime. *Forbes*. https://www.forbes.com/councils/forbesbusinesscouncil/2023/01/31/will-the-dark-warehouse-ever-become-reality-perhaps-not-in-our-lifetime/

Silverman, E. (2019). *How humans and robots work side-by-side in Amazon fulfillment centers.* Phys.Org. https://phys.org/news/2019-06-humans-robots-side-by-side-amazon-fulfillment.html

Simon, M. (2019). Inside the Amazon warehouse where humans and machines become one. *Wired*. https://www.wired.com/story/amazon-warehouse-robots/

Stelmaszak, M., Möhlmann, M., & Sørensen, C. (2024). When algorithms delegate to humans: Exploring human-algorithm interaction at Uber. *MIS Quarterly, 49(1)*, 305-330. https://doi.org/10.25300/MISQ/2024/17911

Stone, P., & Veloso, M. (2000). Multiagent systems: A survey from a machine learning perspective. *Autonomous Robots, 8*, 345-383. https://doi.org/10.1023/A:1008942012299

Tamir, D. I., & Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in Cognitive Sciences, 22*(3), 201-212. https://doi.org/10.1016/j.tics.2017.12.005

Taniguchi, T., Murata, S., Suzuki, M., Ognibene, D., Lanillos, P., Ugur, E., Jamone, L., Nakamura, T., Ciria, A., & Lara, B. (2023). World models and predictive coding for cognitive and developmental robotics: Frontiers and challenges. *Advanced Robotics, 37*(13), 780-806. https://doi.org/10.1080/01691864.2023.2225232

Te'eni, D., Yahav, I., Zagalsky, A., Schwartz, D., Silverman, G., Cohen, D., Mann, Y., & Lewinsky, D. (2023). Reciprocal human-machine learning: A theory and an instantiation for the case of message classification. *Management Science*, Articles in Advance. https://doi.org/10.1287/mnsc.2022.03518

Traeger, M. L., Strohkorb Sebo, S., Jung, M., Scassellati, B., & Christakis, N. A. (2020). Vulnerable robots positively shape human conversational dynamics in a human–robot team. *Proceedings of the National Academy of Sciences, 117*(12), 6370-6375. https://doi.org/10.1073/pnas.1910402117

Ullman-Margalit, E. (1977). Coordination norms and social choice. *Erkenntnis*, 143-155.

Ullmann-Margalit, E. (1977). *The emergence of norms*. Oxford University Press.

Walker, A., & Wooldridge, M. J. (1995). Understanding the emergence of conventions in multi-agent systems. In *Proceedings of the 1st International Conference on Multiagent Systems* (pp. 384-389).

Wawrla, L., Maghazei, O., & Netland, T. (2019). *Applications of drones in warehouse operations. ETH Zurich* (Whitepaper). Available at: https://ethz.ch/content/dam/ethz/special-interest/mtec/pom-dam/documents/Drones%20in%20warehouse%20opeations_POM%20whitepaper%202019_Final.pdf

Weiss, G. (1999). *Multiagent systems: a modern approach to distributed artificial intelligence*. MIT Press.

Wheeler, N. E., Allidina, S., Long, E. U., Schneider, S. P., Haas, I. J., & Cunningham, W. A. (2020). Ideology and predictive processing: coordination, bias, and polarization in socially constrained error minimization. *Current Opinion in Behavioral Sciences, 34*, 192-198. https://doi.org/10.1016/j.cobeha.2020.05.002

Williams, D. (2018). Pragmatism and the predictive mind. *Phenomenology and the Cognitive Sciences, 17*(5), 835-859. https://doi.org/10.1007/s11097-017-9556-5

Wurman, P. R., D'Andrea, R., & Mountz, M. (2008). Coordinating hundreds of cooperative, autonomous vehicles in warehouses. *AI Magazine, 29*(1), 9-9. https://doi.org/10.1609/aimag.v29i1.2082

Yakura, H., Lopez-Lopez, E., Brinkmann, L., Serna, I., Gupta, P., & Rahwan, I. (2024). *Empirical evidence of large language model's influence on human spoken communication* (Working paper). arXiv. https://arxiv.org/abs/2409.01754

Young, A. G., Majchrzak, A., & Kane, G. C. (2021). Organizing workers and machine learning tools for a less oppressive workplace. *International Journal of Information Management, 59*, Article 102353. https://doi.org/10.1016/j.ijinfomgt.2021.102353

About the Authors

Thomas Grisold is a professor of information systems and applied AI at WU Vienna. His research examines how artificial intelligence and other emerging digital technologies transform patterns of organizing. Drawing on his background in cognitive science, he applies interdisciplinary perspectives to the study of artificial intelligence.

Nicholas Berente is the James H. Sweeny III and Alicia Sweeny Collegiate Professor of IT, Analytics, and Operations and the senior associate dean for Academic Programs at Notre Dame's Mendoza College of Business. He received his PhD from Case Western Reserve University and studies digital innovations like artificial intelligence technologies and their organizational, institutional, and ethical implications. He is a senior editor at *MIS Quarterly* and *Information and Organization.*

Stefan Seidel is a professor of information systems at the University of Cologne. He studies how emerging digital technologies—particularly artificial intelligence—are implicated in organizational and institutional change, focusing on digital innovation as well as its governance and regulation. He is a senior editor at *MIS Quarterly.*