Toward Agency in Human–AI Collaboration

Steffen Holter, *ETH Zurich, 8092, Zurich, Switzerland*
Caterina Moruzzi, *University of Edinburgh, EH8 9YL, Edinburgh, U.K.*
Mennatallah El-Assady, *ETH Zurich, 8092, Zurich, Switzerland*

Artificial intelligence (AI) is increasingly evolving from a tool for automating repetitive tasks to an intelligent agent actively engaging in dynamic interactions with humans. As AI becomes more integrated into collaborative contexts, it is essential to examine the factors that shape human–AI interaction. Central to this collaboration is AI agency—the capacity for action and effect—a concept that has remained largely peripheral in existing research. This article addresses this gap by proposing a comprehensive design space for reasoning about agency in human–AI collaboration. We introduce the high-level perspectives of distribution, modeling, and attribution to outline key dimensions that inform the design of agency in such systems. Our methodology combines a literature review with expert interviews to consolidate existing concepts and surface new insights. To exemplify the capacity of our framework, we reason about three mixed-initiative systems through the lens of our conceptual model. Finally, we identify future directions and critical research gaps in this emerging area.

As artificial intelligence (AI) becomes increasingly adept at producing sophisticated responses to diverse user inputs, the focus is shifting toward harnessing these advanced capabilities to their fullest potential. While full automation remains out of reach in most practical applications, involving AI to complement human actions and workflows has yielded some promising results. However, until now, this collaboration has been largely human-centered, with AI reduced to a reactive role. By considering AI agency—the capacity for AI to act and influence outcomes—we can open up new avenues for structuring collaboration that balances initiative between human and AI actors.[1] Attributing a degree of agency to the AI enables it to engage more actively, not only by offering knowledge but also by initiating contextually relevant actions. This shift invites careful reflection on how agency is distributed and negotiated within human–AI systems.

While properties commonly associated with agency—such as interaction and adaptation—have been analyzed at length[2] when it comes to human–AI collaboration, the core concepts of agency and initiative themselves are often relegated to footnotes. Yet, as AI systems assume increasingly prominent roles, it is no longer tenable to sidestep these discussions. A central tension lies in the balance between human control and AI autonomy: users often welcome the convenience of delegating tedious or repetitive tasks to AI but remain hesitant to relinquish decision-making control.[3] This ambivalence underscores the need to critically examine the conditions under which users are willing to embrace the benefits of increased automation while maintaining meaningful oversight.[4] Carefully articulating when, how, and to what extent agency is attributed to the AI is essential for designing collaborative systems that are both effective and aligned with human values.

This article builds on the design space introduced by Holter and El-Assady,[5] which frames human–AI systems through the lenses of agency, interaction, and adaptation. While their work provides a valuable foundation, its treatment of agency is limited in scope and

0272-1716 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.
Digital Object Identifier 10.1109/MCG.2025.3623892
Date of publication 20 October 2025; date of current version 26 January 2026.

Authorized licensed use limited to: ETH BIBLIOTHEK ZURICH. Downloaded on October 13,2026 at 17:07:48 UTC from IEEE Xplore. Restrictions apply.

lacks the depth and flexibility needed to fully account for the complexity of this dimension in collaborative paradigms.

In response, we propose an extended design space specifically aimed at supporting structured reasoning about agency in human–AI collaboration. The conceptual model is intended to provide researchers and system developers with a coherent way of designing agency while also establishing a structured framework through which to discuss this latent property in existing human–AI systems. The design space explores agency from three high-level perspectives—distribution, modeling, and attribution. Each of them is further deconstructed into associated properties describing concrete aspects of human–AI collaboration. The focus is on constraining the discussion to actionable dimensions and complementing them with conceptual considerations where appropriate. The proposed design space was developed through an iterative, interdisciplinary approach that combines a comprehensive literature review with insights gathered from interviews we conducted with expert practitioners across diverse domains.

While our framework is designed to be generalizable and applicable across various human–AI collaborative paradigms, in this article, we ground our discussion primarily within the domain of visual analytics (VA). We have chosen VA as a running example to provide concrete, illustrative cases that demonstrate the framework's utility more clearly. However, the scope and relevance of our proposed dimensions are by no means restricted to this single context, and the design space can readily be applied to other domains involving human–AI collaboration.

In sum, the proposed solution seeks to address three overarching research questions that are central to reasoning about agency in the context of human–AI collaboration.

RQ1) What does it mean to have agency, and what are the key dimensions of agency in human–AI systems?
RQ2) What design tradeoffs arise in systems where agency is shared between human and AI actors?
RQ3) How can agency be distributed and dynamically allocated within a collaborative system?

To further demonstrate the utility of the presented conceptual model, we reflect on agency in three visual mixed-initiative interfaces. Through these use cases, we examine how the dimensions of the design space manifest in existing systems and consider alternative design choices that could shape their agency models differently. Furthermore, we suggest a research agenda and the potential next steps for agency-related discussions in the field of human–AI collaboration.

SHIFT TOWARD AGENCY

Until now, most systems described as supporting human–AI collaboration remain disproportionately human-centered. For instance, in domains, such as VA, the primary emphasis continues to be placed on human-driven analysis, with AI typically limited to a supportive or auxiliary role. However, the emergence of mixed-initiative systems, initially proposed by Horvitz,[6] offers an alternative model characterized by shared agency, where humans and AI actively comanage decisions and interactions.

This shift toward increased AI agency is reflected in recently developed human–AI systems. For example, Podium[7] autonomously refines data rankings by learning user preferences, while ProtoSteer[8] enables collaborative steering of deep learning models through iterative prototype modifications. Other tasks, such as fact-checking, topic modeling, and translation support, have also been supported by active AI participation within analytical workflows.

Surveys of machine learning integration have highlighted how carefully designed AI-driven suggestions, automated analyses, and adaptive interfaces can significantly enhance analytic reasoning, exploration, and decision-making.[9] Related research into coadaptation and guidance deepens this discourse by proposing structured mechanisms through which AI systems learn from and dynamically respond to human input.[10]

Conceptual frameworks have emerged to clarify how agency can be effectively modeled, distributed, and attributed between humans and AI.[11] Yet increasing AI agency introduces critical tensions related to trust, transparency, and balancing users' desire for automation against their hesitation to cede control.[4] Understanding these tensions demands an interdisciplinary approach that incorporates insights from communication, social psychology, and human–computer interaction (HCI).

Despite growing interest in these dynamics, the notion of agency is still too often treated implicitly. While it is embedded in the behavior, functionality, and design rhetoric of many systems, few works engage with agency as a core analytic lens. This leaves a gap in the literature and a missed opportunity for more deliberate, transparent, and critical approaches to designing human–AI collaboration.

Authorized licensed use limited to: ETH BIBLIOTHEK ZURICH. Downloaded on October 13,2026 at 17:07:48 UTC from IEEE Xplore. Restrictions apply.

DEFINING AGENCY

The term "agency" lacks a unified definition in the literature and has a variety of interpretations across domains.[12] Broadly, agency can be understood in two main ways: either as an inherent property of the agent or as something that is granted or attributed to an agent.

In the first interpretation, agency is viewed as an internalized property, something that allows the agent to be in control of their actions and autonomous in performing them.[13,14,15] This understanding of agency is mostly evident in discussions surrounding free will and determinism,[16] as well as in literature that explores the concept of "sense of agency"—the intrinsic awareness of the possession of agency by the subject.[17] Conversely, agency can also be interpreted as a property that is granted to a subject, irrespective of whether the subject inherently possesses it or not. In this context, agency can be seen as a mechanism to determine who takes the initiative for an action and is consequently "in charge" of it. This perspective is discussed in the literature as agency *attribution*, referring to the ascription of the capacity to initiate intentional events to a system or entity.[18]

In this article, we assume this second interpretation of agency. While there exists an open discussion around free will and what it means to have control, adopting a direct action–consequence model is sufficient in the context of human–AI collaboration.[19] Similarly, this work does not delve into the specifics of determinism and probability, nor does it debate the human capacity for intentional behavior. In the human–AI collaboration paradigm, all the participating actors, whether they have agency or not, can be described as agents,[11] and thus, the terms are used interchangeably. Furthermore, our proposed formulation does not distinguish between types of agents (i.e., human or AI), and instead adopts an entirely symmetric structure where no properties are inherently assigned to specific actors. The term mental model is also used for both the internal knowledge representations of humans as well as that of AI agents.

METHODOLOGY

In this section, we outline the two-step approach used to construct the design space. Our methodology combines an interdisciplinary literature review with expert interviews across diverse application domains to both consolidate existing knowledge and introduce new perspectives. Given that agency has largely been addressed only indirectly in the context of human–AI collaboration, the literature review helps anchor our design space in relevant prior work, while the expert interviews enrich it with insights from adjacent fields. This integrative approach helps ensure both conceptual depth and broad applicability. An overview of the process is illustrated in Figure 1.

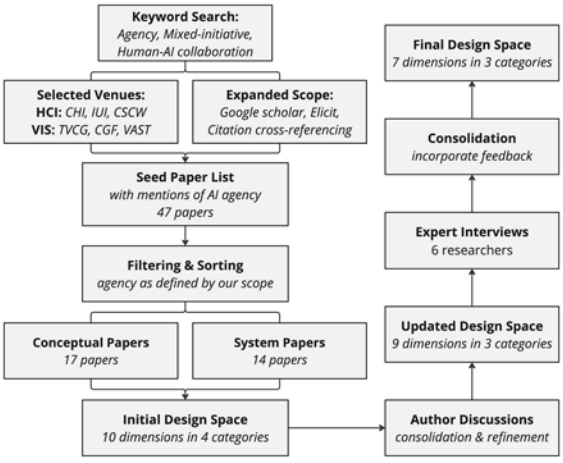

FIGURE 1. Methodology overview: a two-part approach is applied to arrive at the design space for agency. A literature review is used to assemble existing approaches into an initial set of dimensions. In addition, through discussions between authors and domain experts, the design space is refined further into a comprehensive final state.

Literature Review

To generate an initial version of our design space, we systematically surveyed relevant literature to summarize definitions and evaluate related conceptual approaches. As the first step, we performed a keyword search across the proceedings of top-tier VA and HCI venues over the past 10 years. We looked up the terms *agency*, *mixed-initiative*, and *human–AI collaboration* in the titles, abstracts, as well as contents of published works, thus yielding an initial set of potentially relevant papers. Since the notion of *agency* appears in the literature across diverse contexts, we limited our selection to works that explicitly or implicitly discussed forms of AI or machine agency. This determination was made systematically by the authors through a detailed review of each paper's content, as well as supplementary materials, such as publicly available demonstration videos, interactive demos, and source code. However, we intentionally did not restrict our selection based on the specific collaborative tasks described; instead, we aimed to leverage the diversity of tasks present within the literature.

The initial search results reaffirmed that the topic had been considered in a limited capacity in the field

Authorized licensed use limited to: ETH BIBLIOTHEK ZURICH. Downloaded on October 13,2026 at 17:07:48 UTC from IEEE Xplore. Restrictions apply.

of human–AI collaboration and thus we looked to extend our selection scope. We queried "agency" in Google Scholar and cross-referenced relevant citations from the already assembled set of papers. In addition, we made use of an AI-powered research tool *Elicit*[a] to extract any missed mentions of the keywords. This expanded search yielded a broader set of relevant foundational papers from related fields, such as cognitive science, philosophy, and robotics.

Broadly, the initial seed paper list ($n = 47$) comprised two types of published works: *conceptual papers*, which discuss agency as a theoretical concept, and *system papers*, which describe existing mixed-initiative systems in which AI is attributed agency. From this list, we parsed the contents of the papers and narrowed the selection to focus on the relevant scope, excluding papers that presented excessively abstract considerations of the notion of agency. Similarly, work describing existing mixed-initiative systems was only considered relevant if the described AI agents demonstrated capacity for action per our definition. This process yielded a final corpus of papers ($n = 31$). The relatively small body of work that explicitly engages with the idea of agency also further underscores the need for a systematic treatment of the topic.

From this final set, an unstructured initial design space was iteratively constructed by consolidating insights from the conceptual works and ensuring that the dimensions capture the full complexity of the reviewed human–AI systems. This process leveraged the expertise and intuition of the authors to resolve any inconsistencies across works and establish a coherent conceptual model that is generalizable. To account for potential future research perspectives, some additional properties were added to complete the initial set of design space dimensions.

Expert Interviews

To complement the literature review and gauge the perspectives of the wider research community, we conducted a series of semistructured interviews with practitioners and experts in related fields. Here, we describe the properties of the study setup, the participants, and our overall process.

Participants: We interviewed six experts from diverse regions and disciplines, each of whom has actively conducted research on aspects of agency within the past five years. The group included a professor specializing in human–AI collaboration in design, a distinguished professor and international author in digital humanities, and a senior researcher in VA and natural language processing (NLP) at a global technology company. The remaining participants were advanced doctoral students: one focused on multiagent system collaboration, another exploring philosophical considerations of AI-driven scientific discovery, and a third investigating cocreativity.

Study Design: All of the interviews were conducted online in a semistructured manner, with each session lasting around 1 h. To maintain consistency, the discussions were carried out using a *Miro*[b] collaborative sketchboard and a set of pre-prepared guiding questions. Throughout the interviews, participants were encouraged to summarize their main points on post-its while simultaneously explaining their reasoning out loud to the interviewer. This approach allowed us to identify essential arguments, as it required participants to express their complex thoughts in concise and precise language.

Study Process: The interviews were carried out in two stages. The first and primary stage involved an open-ended discussion about the concept of agency within the context of each expert's domain and experience. Our three research questions served as a guide to focus the conversation on aspects of agency most relevant to our research scope. The aim of this stage was to collaboratively explore and surface an unbiased set of insights, perspectives, and conceptual framings. Particular emphasis was placed on exploring the key characteristics that experts deemed as essential for meaningfully describing agency models in different human–AI systems. The second supplementary stage focused on the presentation of our proposed design space and gathering the experts' initial impressions and feedback. In this step, the interviewer explained the design space sequentially and allowed abundant space for any clarifying questions and reflection. Given the limited timeframe, participants naturally focused on the aspects of the model most closely aligned with their areas of expertise. Overall, experts were encouraged to comment on both strengths and limitations, and where appropriate, the interviewer offered counterpoints to promote deeper discussion. All feedback was recorded for later analysis.

DIMENSIONS OF AGENCY

In this section, we outline our design space for agency in human–AI collaboration. Each proposed dimension characterizes a singular aspect of agency in these

[a][Online]. Available: https://elicit.com/

[b][Online]. Available: https://miro.com/

Authorized licensed use limited to: ETH BIBLIOTHEK ZURICH. Downloaded on October 13,2026 at 17:07:48 UTC from IEEE Xplore. Restrictions apply.

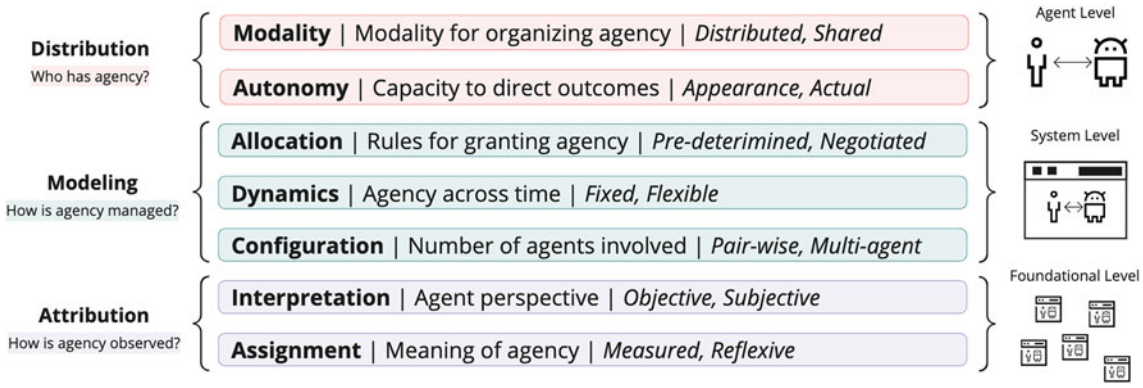

FIGURE 2. Design space for agency in human–AI collaboration features three central categories: (1) distribution, (2) modeling, and (3) attribution. Each perspective analyzes related characteristics of human–AI systems in the form of singular dimensions. In addition, the high-level categories correspond to varying aspects of human–AI collaboration.

systems. For coherence and alignment with our research questions, we group the individual dimensions into three distinct perspectives: distribution (RQ3), modeling (RQ2), and attribution (RQ1). Each perspective expands the scope of the discussion and corresponds to a different level of abstraction for analyzing agency in human–AI collaboration. Specifically, agency can be analyzed on an agent level (distribution), system level (modeling), or foundational level (attribution). While the dimensions are presented in a structured bottom–up order to support readability, they do not imply a strict hierarchy or fixed sequence of analysis. Moreover, this design space specifically applies to systems in which AI agents are capable of taking initiative in analytical tasks. A comprehensive breakdown of the conceptual model and how it relates to human–AI collaboration is summarized in Figure 2.

Our design space is aimed primarily at those designing and conceptualizing human–AI systems. Thus, the focus is on distilling the concrete implementation choices and considerations required to orchestrate this collaboration. Contextual factors, such as the analytic task, user expertise, or problem complexity, inform why one might select a given agency configuration, but they do not directly dictate how to structure agent interactions, define modeling perspectives, or allocate initiative. Consequently, these motivating variables are not included as framework dimensions. Similarly, whether end users explicitly perceive each dimension is not immediately relevant to formulating a clear, actionable conceptual model.

To provide transparency regarding the experts' contributions during the interviews, we have incorporated some of their direct quotes and suggestions into the descriptions of the individual dimensions. The experts are referenced in the text as E1–E6. Furthermore, throughout our analysis, we continuously highlight future research opportunities (RO) in the field, which are then summarized in-depth in the *Research Opportunities* section.

In order to ensure a grounded discussion of the design space, we introduce a running example that will serve as a reference point for reasoning through the proposed dimensions. Let's consider we are tasked with designing a human–AI collaborative framework for a think-tank where humans and AI agents can collectively deliberate potential policy suggestions to tackle climate change. This is an example of a wicked problem[20]—a complex challenge with no definite formulation or answer—and thus yields itself well to an in-depth exploration of agency and intelligent agents. Since some of the discussion exceeds the capabilities of most current human–AI systems, it also acts to represent potential future research directions and challenges in this field.

Agency Distribution

At the root of our design space is the distribution of agency, which characterizes human–AI systems on an agent-to-agent basis. In any collaborative paradigm, the actors are influenced by, or affect, other actors; thus, identifying the agency involved in performing specific actions is nontrivial and involves several dimensions. In other words, the focus is on identifying *who has agency* among the participating actors (RQ3).

Modality | Method for organizing agency

Distributed, Shared

Authorized licensed use limited to: ETH BIBLIOTHEK ZURICH. Downloaded on October 13,2026 at 17:07:48 UTC from IEEE Xplore. Restrictions apply.

Agency can be split among actors according to various modalities that are distinguished on the basis of the intentions of the agents.[21] *Distributed* agency entails that the actions performed by agents combine and lead to a single outcome. These actions may also converge by chance; however, the intention from the side of the agents to act in coordination is not necessary.[22] Every agent is independent in performing the action and responsible for the action performed. Alternatively, for *shared* agency, the action is performed together by the agents. It doesn't converge by chance and instead originates from the intention of the agents involved to act in coordination.[23,24] Not all actors need to be equal contributors but each actor needs to be responsive to the intentions of the others. In our example, the climate change discussion would benefit from *shared* agency as the actions (e.g., policy decisions) call for active coordination and alignment on intentions between actors.

Autonomy | Capacity to make choices

Appearance, Actual

The notion of autonomy is a prerequisite for the strategies discussed above. For an actor to assume agency, it needs to have the capacity to proactively make choices and manage the direction of outcomes.[19,25,26] Yet, it is often difficult to distinguish between the *appearance* of autonomy (simulacrum) and *actual* autonomy. For example, the most common modes of AI output are synthesis and simulation, both of which may be attributed to only an appearance of autonomy. It also follows that agents generally cannot be held accountable for their actions unless they are acting autonomously[27,28]; however, the exact nuance of this relationship assumes a more inherent definition of agency RO4. Experts contributed varying insights as E6 remained adamant that "intention and consequences are essential to attribute agency" while E4 disagreed and stated that "intentionality is more related to planning and deliberation than the act of taking action itself." For the agents in our climate debate scenario, they can be assumed to possess *actual* autonomy to guarantee proactive engagement in the collaboration and ensure policy choices are made by all involved parties.

Agency Modeling

Modeling agency generalizes the discourse by one level and looks at human–AI dynamics on a per-system basis and how this changes across interactions between the participating actors. This entails evaluating the conditions by which agency is exchanged and how that changes across time. As such, the focus is on deciding *how the agency is managed* and what are the general conceptual design considerations that need to be accounted for when introducing intelligent agents into human–AI systems (RQ2).

Allocation | Rules for granting agency

Predetermined, Negotiated

A fundamental design consideration for human–AI systems is how agency is assigned to different actors. E3 described this dimension as "particularly crucial when thinking about how to design new types of systems in the future." As proposed by Holter and El-Assady,[5] the allocation process can be modeled in two general ways. A *predetermined* approach relies on a defined structure orchestrated by the system itself, completely exempting the individual agents from the decision. Alternatively, the agency can be decided on a *negotiation* basis in which the system only provides general conditions for the allocation, but the decision is made by the participating actors in accordance with some attribution approach.[23] As pointed out by E1, this type of opportunistic negotiation relies on the existence of a deterministic way of collectively deciding agency that avoids conflicting objectives and subjectivity RO3. Returning to our running example, ideally, the actor that is best suited for a particular policy topic should be able to assume agency through the process of *negotiation*.

Dynamics | Agency across time

Fixed, Flexible

Systems involving repeated interaction inherently introduce a temporal component to managing agency.[29] The length of time for which the agency is allocated after each exchange and when it can be transferred again dictates the modeling process. A *fixed* method creates structure along the time axis by designating trigger points when initiative can transition between agents. By extension, the exact period for which an actor is in control is also specified. Removing such constraints and introducing a fully *flexible* strategy would allow agents to shift agency without limitation. However, such a

Authorized licensed use limited to: ETH BIBLIOTHEK ZURICH. Downloaded on October 13,2026 at 17:07:48 UTC from IEEE Xplore. Restrictions apply.

free-form approach might still be constrained by the specific analytical task at the center of the collaboration process. For a contentious problem, such as policy debate, the balance of human and AI agency should not remain static but should instead evolve dynamically as the collaboration progresses. Employing a predominantly *flexible* agency model is beneficial, as it allows initiative to shift fluidly between agents, adapting naturally to changing circumstances and emerging insights throughout the interaction.

Configuration | Number of agents involved

Pairwise, Multiagent

In addition, the modeling process is directly influenced by the number of actors among whom agency must be distributed. The usual formulations of agency in mixed-initiative systems assume a *pairwise* model where a singular human and AI are competing for agency. Such an approximation is often adopted in the design of practical applications to simplify the modeling process. Ideally, larger scale *multiagent* human–AI systems should incorporate a way to manage agency among a number of intelligent agents. However, this adds to the challenges of initiative allocation as scaling up is nontrivial. For example, compared to a pairwise approach, identifying the optimal recipient of agency from a plethora of potential candidate agents is far more nuanced.[30] As suggested by E2, these multiagent configurations also "require thinking about hierarchies between agents in shared agency contexts." Yet, for our running example, the solving of this multiobjective problem will require a range of perspectives and thus will be a multiagent collaboration with various human and AI actors involved.

Agency Attribution

Shifting to the highest level of abstraction, the question of attribution examines agency as a fundamental concept and pertains to all systems related to the human–AI collaboration paradigm. This set of dimensions functions as a way to reduce ambiguity and allow for structured discourse around different forms of agency. The fundamental question at the center of this category is *how is agency observed* and how can this be explained (RQ1).

Interpretation | Agent perspective

Objective, Subjective

Any discussion of the attribution of agency needs to address the issue of interpretation and agent perspective. The dynamics between participating actors are shaped by the expectations each actor has with regard to the intentions, capabilities, and goals of the other agents RO2. Thus, the interpretation of agency can be influenced by the internal mental model of the agent, which is also subject to change across repeated interaction.[31,32] Actors can adopt this type of *subjective* interpretation of agency when they observe changes at the level of their own mental model. Alternatively, agents can adopt an *objective* interpretation of agency when they evaluate the outcomes and effects of agency by observing changes that occur in the external environment. E4 wondered if "this dimension could also equally be called *internal* and *external*" and warned, "it is often hard to make this distinction because of how external stimulus impacts internal perceptions." Understanding these perspectives is key to analyzing how agents form beliefs about their own agency and that of others. In the design of the example system, a *subjective* perspective may emerge due to the succession of discussions and interactions that contribute to the eventual policy decision. Each actor may perceive themselves as having played a decisive role, even if they did not directly generate the final outcome.

Assignment | Meaning of agency

Measured, Reflexive

The absence of a unified definition of agency in human–AI systems makes it essential to establish clear criteria for determining when an actor can be said to possess agency.[22,33] A direct and *measurable* classification of agency relies on quantifiable effects on the analytic task or environment. In other words, agency can be determined by comparing the different actors' capacity to solve the objective and the direct impact they have on the task itself. Alternatively, agency can be understood in *reflexive* terms, encompassing both an agent's internal sense of initiative and their interpretation of agency in others. This includes self-perception, where actors assess their own influence and reflect on their inner sense of initiative and control. It is comparable to a sense of self-awareness and opinion regarding contributions made toward a goal.[15] In addition, this definition also includes attributed perception, where actors assign agency to others based on observed behaviors or assumptions regarding their intent. In both cases, agency is constructed through reflection, rather than direct measurement.

Authorized licensed use limited to: ETH BIBLIOTHEK ZURICH. Downloaded on October 13,2026 at 17:07:48 UTC from IEEE Xplore. Restrictions apply.

This dimension, therefore, addresses the basis on which agency is identified or assigned. It defines what is considered agency within a given system. In the policy deliberation scenario, agency is attributed based on which actor takes action, such as proposing a policy or contributing meaningfully to the process. This constitutes a case of *measured* assignment.

USE CASES

In this section, we apply the proposed design space to three use cases in VA featuring mixed-initiative interfaces. Using our conceptual model, we examine how AI agency manifests within each of the systems. This analysis serves to illustrate both the practical utility of the design space and its value in facilitating structured discussions around agency in human–AI collaboration.

As previously noted, many existing systems do not address the full spectrum of agency dimensions outlined in this work. However, analyzing real-world use cases allows us to assess the current applicability of the framework and identify opportunities for refinement. In each case, we focus the discussion on a subset of relevant dimensions, selected based on contextual relevance and the nature of the system. Where applicable, we also reflect on how each system could be modified to support alternative agency models according to our conceptual framework. All selected interfaces exhibit some form of AI agency, ensuring their relevance to our research scope.

The systems selected for our use cases were chosen systematically based on three primary criteria: 1) diversity of application domains, ensuring broad relevance across different contexts; 2) variety in human–AI agency configurations, clearly demonstrating distinct design choices within our proposed framework; and 3) sufficient documentation and notability, providing detailed and reliable information for thorough analysis. The chosen systems are shown in Figure 3.

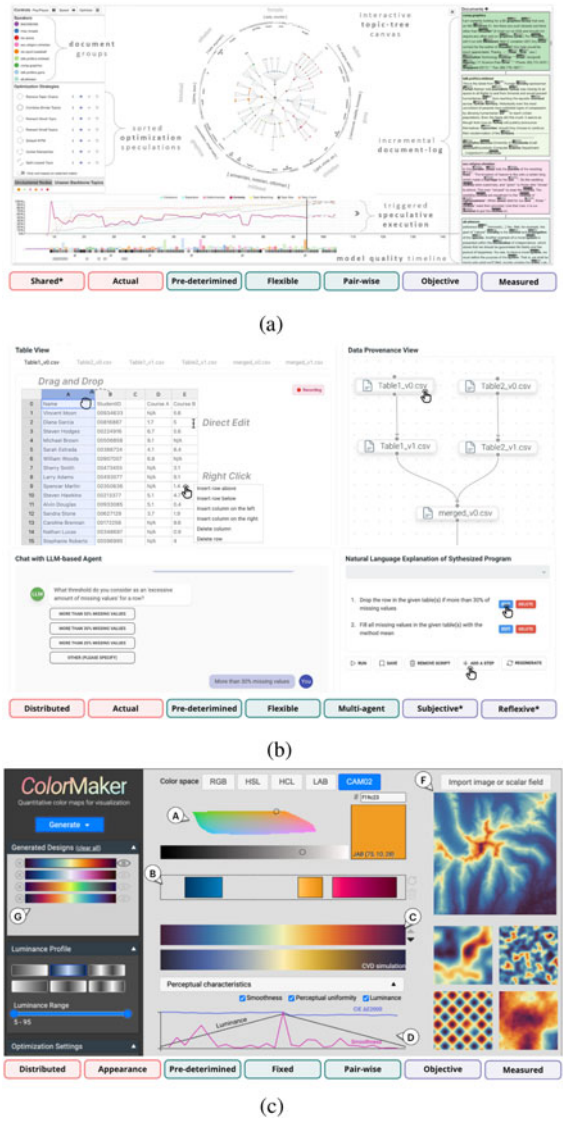

FIGURE 3. Overview of the three mixed-initiative visual interfaces discussed in the use cases, including their characterization based on the proposed design space. (a) SpecEx: User steerable speculative execution.[34] (b) Dango: Data wrangling system using LLMs.[35] (c) Color Maker: Creating accessible color maps.[36]

Building Topic Models

In this first use case, we examine how agency is exhibited in the mixed-initiative VA framework proposed by El-Assady et al.[34] for speculative execution (Figure 3a). The interface incorporates user-steerable preview mechanisms into the decision-making process for a topic modeling task.

The system primarily follows a *distributed* agency strategy, wherein the human user assumes control only when the AI's internal model metrics indicate high uncertainty. However, from a task-centric perspective, the system may also be interpreted as employing a *shared* strategy, with topic modeling optimization resulting from the combined efforts of both agents. This dual interpretation underscores a potential design opportunity, as certain system properties may benefit from more flexible mappings across agency dimensions RO1. The model's individual optimization choices reflect *actual* autonomy, while the overall agency model is *pairwise* and *predetermined*, with the system delegating complex cases to the

Authorized licensed use limited to: ETH BIBLIOTHEK ZURICH. Downloaded on October 13,2026 at 17:07:48 UTC from IEEE Xplore. Restrictions apply.

human while the AI retains control over the remaining optimization process. In contrast, other systems for topic modeling, such as that of Sperrle et al.[30] distribute decision-making among multiple agents, thereby exemplifying *multiagent* and *negotiated* agency models. In terms of temporal dynamics, *SpecEx* exhibits a *flexible* approach, as agency is transferred based on internal uncertainty estimates and user-triggered toggles. The acting agent is determined by observing direct effects on the environment—specifically, refinement actions taken with respect to the task. This results in *objective* and *measured* attribution.

Collaborative Data Wrangling

As the next example, we reflect on the Dango[35] data wrangling tool, where a nonexpert user collaborates with multiple AI agents to generate scripts for processing and transforming raw data (Figure 3b). The system focuses on enhancing communication between large language model (LLM) agents and users, facilitated through direct demonstration and natural-language interactions. Dango is designed as a *multiagent system*, incorporating separate agents for analysis, synthesis, and explanation, each complementing the human user. These agents work in unison to accomplish the overarching task of synthesizing data wrangling scripts.

Overall, the tool employs a *distributed* agency model, where each agent is responsible for a distinct step in the workflow. For example, the analysis agent is dedicated solely to inferring user intent from demonstrations or language input and does not participate in the actual generation of scripts. The roles of the agents are *predetermined* and operate according to *fixed* interaction phases. However, the perceived locus of agency can shift depending on the attribution model applied. From a *measured* and *objective* perspective, agency is attributed to the synthesis agent, which takes initiative by generating the scripts. While the human may attempt to guide or validate the outcome, the direct manipulation of the pipeline's output is the sole responsibility of this agent. In contrast, a *reflexive* and *subjective* view might assign agency to the human user, whose perceived intent is central to the modeling process. This difference in perspective also informs which agents possess *actual* autonomy and which merely exhibit the *appearance* of agency.

This use case also illustrates that identifying who holds control in the decision-making process is not always straightforward. As such, selecting a clear attribution criterion is essential for determining how agency is assigned within a system. Developing a concrete vocabulary, as outlined in our design space, further supports the communication of design choices in mixed-initiative systems.

Creating Color Maps

In this final system,[36] a human and an AI agent engage in a mixed-initiative manner to design color maps (Figure 3c). The task is complicated by the need to balance perceptual standards with the personal preferences of users. The agents do not act in coordination and adopt a *distributed* agency model where the AI incrementally works around provided specifications. The agents only have an *appearance* of autonomy and act according to a clear objective function to complement user inputs by performing tasks, such as meticulously positioning control points to achieve a smooth and uniform color gradient. This reflects a broader trend in tools for nonexpert users, where systems often provide the *appearance* of autonomy and user control while automating complex domain-specific decisions behind the scenes. The agency model is *predetermined* and follows *fixed* dynamics for the *pairwise* interaction. Depending on the preferences provided by the user, the AI is responsible for the rest of the design choices. It represents an interactive, user-driven optimization, where the AI's role is automating the majority of the color map generation process. It makes sense to view the agency attribution as *measured* and *objective* as the agents' contribution on the resulting visual color map signifies who has initiative at each step. This use case also acts as an example of the most typical configuration of agency that we encountered in our literature review. Systems that leverage AI agency and automation in very specific contexts to complement human user interaction.

APPLYING THE DESIGN SPACE

The described design space functions primarily as a conceptual framework for discussing agency in current and future collaboration paradigms. However, it also has the potential to offer concrete, actionable tools for developers building human–AI systems across various domains.

The design space represents a unified vocabulary for designers and stakeholders to reason about human and AI roles in these complex settings, closing the communication gap that has long hindered discussions of agency. This shared language enables teams to articulate critical design decisions, such as how to determine whether a human or an AI agent initiates a given analysis step, and ensures that essential considerations, such as modality and assignment, receive

Authorized licensed use limited to: ETH BIBLIOTHEK ZURICH. Downloaded on October 13,2026 at 17:07:48 UTC from IEEE Xplore. Restrictions apply.

due attention. Similarly, it helps reduce ambiguity and misalignment stemming from differing notions of what agency can and should entail.

Furthermore, the individual dimensions function as a systematic checklist during both design and evaluation. By assessing a system against each dimension, developers can identify imbalances or omissions early in the process. For example, they may discover that the agency model of a planned tool is too static and does not accommodate the evolving nature of collaboration that extends over multiple sessions. Detecting such issues in advance supports iterative refinement and helps avoid revisions in the late stages.

Finally and most importantly, the framework stimulates innovation by revealing underexplored combinations of design choices. It invites designers to imagine new configurations, for instance, a multiagent system employing negotiated agency, and then use the dimensions to prototype and evaluate these possibilities. It also facilitates comparison with other systems employing similar agency configurations, enabling designers to draw insights and inspiration for their own designs.

Ultimately, our design space is more than a descriptive taxonomy. It is a practical blueprint that empowers system developers to make informed decisions, communicate tradeoffs clearly, and integrate agentic capabilities systematically.

GENERALIZABILITY

The general purpose framing of our proposed design space assumes a deliberately broad scope. Our aim is to capture overarching patterns that recur across diverse human–AI contexts and to establish a widely useful conceptual foundation. To validate the relevance of these guidelines, we also examine them across a variety of human–AI systems and tasks. While this level of abstraction supports broad applicability, we also recognize that specialized fields may introduce unique constraints and affordances that can significantly shape agency models.

For this reason, the design space should not be considered in an isolated setting but instead in tandem with other established frameworks. When additional dimensions are needed to reflect the specifics of a given human–AI collaboration scenario, they can be integrated from existing domain-focused models. This supplementation enhances applicability for practitioners by preserving the descriptive generalization of the design space while providing a structure that supports the instantiation of more concrete practical examples. For instance, in VA, Schulz et al.[37] proposed a framework that describes VA tasks, such as exploratory analysis, confirmatory analysis, and presentation. Mapping such task types onto our dimensions may potentially reveal actionable patterns: exploratory analysis could benefit from flexible and negotiated agency, whereas confirmatory analysis may require more fixed and predetermined allocations. This illustrates how different task demands can imply different rules for agency, but also how they can still be characterized with our model.

Although specifying exact correspondences between domain-specific frameworks and our general model lies beyond the scope of this article, these connections can be made by practitioners themselves. We look to provide the building blocks that domain experts can use to derive application-dependent design subspaces. This combined approach serves as a compromise that ensures context-specific practical impact while retaining the conceptual foundation needed to compare and reason across domains.

RESEARCH OPPORTUNITIES

The proposed conceptual model represents a first step in reasoning about agency in human–AI systems; however, we acknowledge that there are still limitations to be addressed in future work. In the following, we highlight a prospective research agenda and five opportunities as a call for action to the research community.

RO1) Isolating Dimensions: The individual dimensions that comprise the design space are not perfectly orthogonal, and there is a lot of contextual specificity that is difficult to estimate using our high-level approach. As mentioned by E4 and E5, especially how agency is observed can be hard to disentangle in some contexts. Thus, there is a need to develop frameworks to systematically isolate the individual characteristics of agency proposed in this work. Creating a playground environment where users can configure various agency models and observe the resulting phenomena with respect to a fixed analytical task would allow gaining invaluable insight into the practical applications of agency. Furthermore, it became evident when attempting to classify existing systems as use cases that the correct level of granularity for each dimension can be difficult to estimate and by design requires making a tradeoff. While overall the design space is intended to cover the entire scope of human–AI collaboration, there are situations where a system property might not align perfectly with the binary or trinary values proposed by our model and would benefit from more specific characterization. However,

Authorized licensed use limited to: ETH BIBLIOTHEK ZURICH. Downloaded on October 13,2026 at 17:07:48 UTC from IEEE Xplore. Restrictions apply.

foregoing some flexibility is a necessary tradeoff to maintain enough generalizability to find commonality between similar systems.

RO2) Agent Communication: As described in the design space, agency is often dependent on the mental models of the participating human and AI actors who are solving the collaborative analytical task. Therefore, how and according to what criteria relevant information is communicated between the agents is a fundamental consideration for any mixed-initiative system. This need to share information also extends to agents adapting and learning over the course of the task, as it impacts the allocation of agency. Future work should look at shared information representation and explicitly evaluate the conditions for successful communication between different actors. For example, human-to-human or AI-to-AI interaction clearly has different considerations than human–AI communication. All of these relationships need to be outlined and related to the model of agency.

RO3) Negotiating Agency: The conceptual model proposed negotiation as a way to allocate agency; however, this premise carries a lot of subtlety that needs to be explored further. For example, the exact mechanisms and conditions for the negotiation should be imposed by the system itself in the form of structured rules. This begs the question: are the agents aligned on their goals and negotiating to optimize the task performance, or do agents have individual ambitions (e.g., human actors trying to reduce their workload)? Similarly, since many relevant tasks lack an objective ground truth, resolving conflict between opposing perspectives becomes nontrivial. E1 and E2 added that in multiagent systems, these challenges become amplified and require the introduction of hierarchies or similar rule sets to facilitate the negotiation. A further layer of complexity arises when considering how the negotiation process evolves along the temporal dimension. In particular, users often engage in long-term, progressive collaboration with AI, necessitating continuous adjustment and renegotiation of roles throughout ongoing and evolving tasks.

RO4) Accountability: On a conceptual level the design of agency in human–AI systems also needs to address risk and accountability.[38] By shifting decision-making control to various AI agents, it becomes immediately unclear who assumes responsibility for the actions taken. Intuitively, only actors deemed as possessing *actual* autonomy should be viewed as being responsible for their actions. Yet, it can be reasoned that the mechanisms of the allocation process should also share liability. Thus, the risk of decisions and establishing accountability needs to be thoroughly examined as a prerequisite for future agency discussions, especially in real-domain applications. This premise also underlies the current tension in mixed-initiative settings between outsourcing tasks and reluctance to yield control of key decisions.

RO5) Human-Centered Evaluation: As the conceptual model is centered around collaboration, it is important to also systematically evaluate the human perspective. Performing user studies with various types of agency models would provide invaluable insight for the design of future human–AI systems. For example, while we can postulate that more AI agency results in user empowerment through increased efficiency, this still needs to be confirmed and quantified. Furthermore, the design of mixed-initiative frameworks would benefit from a unified set of metrics that describe a successful system. Some properties used to evaluate machine learning models (e.g., quality, transparency, trustworthiness, etc.) generalize to human–AI systems but there are also additional parameters that need to be considered. For example, reasoning about capacity tradeoffs and optimally making use of human expertise in these collaboration paradigms.

CONCLUSION

In this article, we introduce a conceptual model for agency in human–AI collaboration. This work represents an initial effort to bring structure to the complex and often implicit notion of AI agency, providing experts with a foundation for reasoning about mixed-initiative systems in a more comprehensive way. VA, in particular, is increasingly embracing AI as an active partner in the analytical process, underscoring the importance of understanding how agency is shared, negotiated, and interpreted. Our proposed design space is structured across three levels of abstraction—*distribution*, *modeling*, and *attribution*—each consisting of distinct dimensions characterizing agency. Through this initial conceptualization, we aim to encourage continued exploration and a more deliberate research agenda focused on the design, evaluation, and ethical considerations of agency in human–AI collaboration.

ACKNOWLEDGMENTS

The work of Steffen Holter was supported by the Swiss National Science Foundation (SNSF) within the Project 10003068. The work of Caterina Moruzzi was supported by the Arts and Humanities Research Council through Bridging Responsible AI Divides (BRAID) programme under Grant AH/X007146/1.

Authorized licensed use limited to: ETH BIBLIOTHEK ZURICH. Downloaded on October 13,2026 at 17:07:48 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

1. A. Satyanarayan and G. M. Jones, "Intelligence as agency: Evaluating the capacity of generative AI to empower or constrain human action," An MIT Exploration of Generative AI, 2024. [Online]. Available: https://vis.csail.mit.edu/pubs/intelligence-as-agency
2. S. Amershi et al., "Guidelines for human-AI interaction," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–13.
3. D. Bennett, O. Metatla, A. Roudaut, and E. D. Mekler, "How does HCI understand human agency and autonomy?," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2023, pp. 1–18.
4. J. Heer, "Agency plus automation: Designing artificial intelligence into interactive systems," *Proc. Nat. Acad. Sci.*, vol. 116, no. 6, pp. 1844–1850, 2019.
5. S. Holter and M. El-Assady, "Deconstructing human-AI collaboration: Agency, interaction, and adaptation," *Comput. Graphics Forum*, vol. 43, no. 3, 2024, Art. no. e15107.
6. E. Horvitz, "Principles of mixed-initiative user interfaces," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 1999, pp. 159–166.
7. E. Wall, S. Das, R. Chawla, B. Kalidindi, E. T. Brown, and A. Endert, "Podium: Ranking data using mixed-initiative visual analytics," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 1, pp. 288–297, Jan. 2018.
8. Y. Ming, P. Xu, F. Cheng, H. Qu, and L. Ren, "ProtoSteer: Steering deep sequence model with prototypes," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 1, pp. 238–248, Jan. 2020.
9. Q. Wang, Z. Chen, Y. Wang, and H. Qu, "A survey on ML4VIS: Applying machine learning advances to data visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 12, pp. 5134–5153, Dec. 2022.
10. I. Pérez-Messina, D. Ceneda, M. El-Assady, S. Miksch, and F. Sperrle, "A typology of guidance tasks in mixed-initiative visual analytics environments," *Comput. Graphics Forum*, vol. 41, no. 3, pp. 465–476, 2022.
11. S. Monadjemi, M. Guo, D. Gotz, R. Garnett, and A. Ottley, "Human-computer collaboration for visual analytics: An agent-based framework," *Comput. Graphics Forum*, vol. 42, no. 3, pp. 199–210, 2023.
12. C. Moruzzi, "Creative agents: Rethinking agency and creativity in human and artificial systems," *J. Aesthetics Phenomenol.*, vol. 9, pp. 245–268, 2022.
13. G. Andrada, R. W. Clowes, and P. R. Smart, "Varieties of transparency: Exploring agency within AI systems," *AI Soc.*, vol. 38, no. 4, pp. 1321–1331, 2023.
14. V. Chambon, N. Sidarus, and P. Haggard, "From action intentions to action effects: How does the sense of agency come about?," *Front. Hum. Neurosci.*, vol. 8, 2014, Art. no. 320.
15. R. Legaspi, Z. He, and T. Toyoizumi, "Synthetic agency: Sense of agency in artificial intelligence," *Curr. Opin. Behav. Sci.*, vol. 29, pp. 84–90, 2019.
16. R. Kane, *The Oxford Handbook of Free Will*. Oxford, U.K. : Oxford Univ. Press, 2011.
17. M. Pagliari, V. Chambon, and B. Berberian, "What is new with artificial intelligence? Human–agent interactions through the lens of social agency," *Front. Psychol.*, vol. 13, 2022, Art. no. 954444.
18. S. Srinivasa and J. Deshmukh, "The evolution of computational agency," in *Novel Approaches to Information Systems Design*. Palmdale, PA, USA: IGI Global Scientific, 2020, pp. 1–19.
19. R. L. Walker, "Respect for rational autonomy," *Kennedy Inst. Ethics J.*, vol. 19, no. 4, pp. 339–366, 2009.
20. B. W. Head and J. Alford, "Wicked problems: Implications for public policy and management," *Admin. Soc.*, vol. 47, no. 6, pp. 711–739, 2015.
21. R. Tuomela, "We-intentions revisited," *Philos. Stud.*, vol. 125, pp. 327–369, 2005.
22. P. Bisconti, A. McIntyre, and F. Russo, "Synthetic socio-technical systems: Poiêsis as meaning making," *Philosophy Technol.*, vol. 37, no. 3, 2024, Art. no. 94.
23. H. Kang and C. Lou, "AI agency vs. human agency: Understanding human–AI interactions on Tiktok and their implications for user engagement," *J. Comput.-Mediated Commun.*, vol. 27, no. 5, Art. no. 14, 2022.
24. S. Nyholm, "Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci," *Sci. Eng. Ethics*, vol. 24, no. 4, pp. 1201–1219, 2018.
25. A. Sharma, S. Rao, C. Brockett, A. Malhotra, N. Jojic, and B. Dolan, "Investigating agency of LLMs in human-AI collaboration tasks," in *Proc. 18th Conf. Eur. Chapter Assoc. Comput. Linguistics*, Mar. 2024, pp. 1968–1987.
26. M. Luck and M. d'Inverno, "A formal framework for agency and autonomy," in *Proc. 1st Int. Conf. Multi-Agent Syst.*, 1995, pp. 254–260.
27. J. Symons and S. Abumusab, "Social agency for artifacts: Chatbots and the ethics of artificial intelligence," *Digit. Soc.*, vol. 3, no. 1, 2024, Art. no. 2.
28. M. E. Bratman, *Shared Agency: A Planning Theory of Acting Together*. Oxford, U.K.: Oxford Univ. Press, 2013.
29. G. Yu and A. Blackwell, "Effects of timing on users' agency during mixed-initiative interaction," in *Proc. 31st Brit. Comput. Soc. Hum. Comput. Interaction Conf.*, 2017, pp. 1–12.

Authorized licensed use limited to: ETH BIBLIOTHEK ZURICH. Downloaded on October 13,2026 at 17:07:48 UTC from IEEE Xplore. Restrictions apply.

30. F. Sperrle, H. Schäfer, D. Keim, and M. El-Assady, "Learning contextualized user preferences for co-adaptive guidance in mixed-initiative topic model refinement," *Comput. Graphics Forum*, vol. 40, no. 3, pp. 215–226, 2021.
31. T. Kulesza, S. Stumpf, M. Burnett, and I. Kwan, "Tell me more?: The effects of mental model soundness on personalizing an intelligent agent," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, May 2012, pp. 1–10.
32. M. Langer et al., "What do we want from explainable artificial intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research," *Artif. Intell.*, vol. 296, 2021, Art. no. 103473.
33. M. van Lier, "Introducing a four-fold way to conceptualize artificial agency," *Synthese*, vol. 201, no. 3, 2023, Art. no. 85.
34. M. El-Assady, F. Sperrle, O. Deussen, D. Keim, and C. Collins, "Visual analytics for topic model optimization based on user-steerable speculative execution," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 1, pp. 374–384, Jan. 2019.
35. W.-H. Chen, W. Tong, P. Case, Amanda, and T. Zhang, "Dango: A mixed-initiative data wrangling system using large language model," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2025, pp. 1–28.
36. A. Salvi, K. Lu, M. E. Papka, Y. Wang, and K. Reda, "ColorMaker: A mixed-initiative approach to creating accessible color maps," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2024, pp. 1–17.
37. H.-J. Schulz, T. Nocke, M. Heitzler, and H. Schumann, "A design space of visualization tasks," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 12, pp. 2366–2375, Dec. 2013.
38. N. J. Enfield and P. Kockelman, *Distributed Agency*. Oxford, U.K.: Oxford Univ. Press, 2017.

STEFFEN HOLTER is working toward the doctoral degree at the IVIA Lab, ETH Zurich, Zurich, 8092, Switzerland. His research interests include agentic human–AI collaborative workflows and optimally leveraging the strengths of human and AI agents to achieve superior joint performance. Holter received his master's degree in computing from Imperial College London, London, U.K. He is the corresponding author of this article. Contact him at sholter@ethz.ch.

CATERINA MORUZZI is a Chancellor's fellow at the Institute for Design Informatics, University of Edinburgh, Edinburgh, EH8 9YL, U.K. She leads the Creativity, AI, and the Human research cluster at the Edinburgh Futures Institute and is a member of the Inclusive AI Lab, University of Utrecht, and of the C2PA research consortium. Her research interests include computational creativity, the philosophy of artificial intelligence, and human–computer interaction. Moruzzi received her Ph.D. degree in philosophy from the University of Nottingham, Nottingham, U.K. Contact her at cmoruzzi@ed.ac.uk.

MENNATALLAH EL-ASSADY is currently an assistant professor at ETH Zurich, Zurich, 8092, Switzerland, where she leads the Interactive Visualization and Intelligence Augmentation Lab. She works at the intersection of data analysis, visualization, computational linguistics, and explainable artificial intelligence. Her research focuses on combining data mining and machine learning techniques with visual analytics, specifically for text data. Contact her at menna.elassady@ai.ethz.ch.

Authorized licensed use limited to: ETH BIBLIOTHEK ZURICH. Downloaded on October 13,2026 at 17:07:48 UTC from IEEE Xplore. Restrictions apply.