Which Biases and Reasoning Pitfalls Do Explanations Trigger? Decomposing Communication Processes in Human–AI Interaction

Mennatallah El-Assady, *ETH AI Center, CH-8092, Zurich, Switzerland*

Caterina Moruzzi, *University of Konstanz, DE-78467, Konstanz, Germany*

Collaborative human–AI problem-solving and decision making rely on effective communications between both agents. Such communication processes comprise explanations and interactions between a sender and a receiver. Investigating these dynamics is crucial to avoid miscommunication problems. Hence, in this article, we propose a communication dynamics model, examining the impact of the sender's explanation intention and strategy on the receiver's perception of explanation effects. We further present potential biases and reasoning pitfalls with the aim of contributing to the design of hybrid intelligence systems. Finally, we propose six desiderata for human-centered explainable AI and discuss future research opportunities.

Mixed-initiative systems have been successfully integrated in multiple domain applications, where human and artificial intelligence (AI) augment one another. To achieve such *hybrid intelligence*,[1] communication interfaces (e.g., interactive visual analytics workspaces) are essential. To facilitate the analysis through such interfaces, tailored interaction workflows are studied and researched. Generally, the goal of interactive human–AI collaboration for hybrid intelligence is to perform efficient and effective problem-solving and decision making. Hence, one of the main principles of mixed-initiative systems is "*Minimal Feedback for Maximal Gain,*" i.e., involving the human or the AI agent where their intelligence is most effective.

As depicted in Figure 1, this leads to a spectrum of mixed-initiative workflow designs, ranging from manual tasks with an AI in the loop to automatic tasks with humans in the loop. Systems that are designed within this spectrum typically perform multiobjective optimizations. They further need to balance the degree of automation versus manual work based on the tradeoff between costs and risks of the task; the data ambiguity and contextualization; and the subjectivity and personalization degree of the analysis. Thus, by balancing these aspects, tailored interfaces can allow for human and AI to effortlessly augment each other, giving humans a superpower through the logic, scalability, and computing power of the AI, and in turn, giving AIs a superpower through the perception, creativity, and general world knowledge of humans.

Especially interactive visual analytics techniques can empower humans through different, effective communication-support techniques.[24]

These interactions allow both agents to communicate their problem-solving rationales and explain their decision making, filling each other's *knowledge gaps*.[13] To explain the inner working of AI models, explainable AI (XAI) models have been researched and developed. In general, the processes of interactive and explainable AI encompass three stages[23]: 1) *Understanding* of the AI's decisions and behavior; 2) *Diagnosis* of the AI's performance and applicability; and 3) *Refinement* of the AI models for the given users, tasks, and data. Explanations follow different strategies (e.g., inductive, deductive, contrastive) and use different mediums for communication (e.g., visualization or verbalization).

However, achieving meaningful and effortless communication through explanations is challenging. Human explanations are usually contrastive, selective, social, and adaptive.[16] They can be used for education

0272-1716 © 2022 EU
Digital Object Identifier 10.1109/MCG.2022.3200328
Date of publication 12 September 2022; date of current version 13 December 2022.

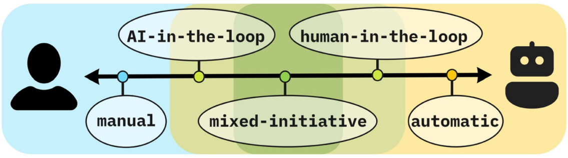


FIGURE 1. Intelligence augmentation spectrum.

(learning and teaching), the presentation of alternative opinions and information, or persuasion based on a belief system. Depending on the communication modality and dynamic, humans understand each other's explanations as suggestions, facts, or decisions.[13] This allows them to assess the associated knowledge content and avoid miscommunication.

Hence, based on the observation of human–human communication, open questions in mixed-initiative research are *how does communication between humans and AIs differ in their dynamics,* and *which challenges arise when explaining problem-solving and decision-making processes?*

As depicted in Figure 2, one of the fundamental challenges is that humans communicating with AIs will presuppose notions of explanation and interaction that are based on their experience communicating with fellow humans. On the other hand, an AI also has predetermined notions about explanation and interaction. For example, humans contextualize and adapt their explanations and interactions dynamically based on the sociotechnical context in which they occur.[3] Hence, they expect others to use similar conceptual reference points in their communication. However, AIs are typically designed to be general-purpose applications with less nuance in contextualization or adaptation of their communication.

Communication is conditional, contextual, and time dependent. Hence, studying the impact of communication dynamics on both agents, the *sender and receiver,* as well as on both humans and AIs is crucial to avoid miscommunication problems. This is especially important to assess potential biases and reasoning pitfalls that can be triggered through explanations and interactions. Such issues have not yet been studied extensively in XAI research. However, as the reach of AI-based applications scales to many users and scenarios, we crucially need to make decisions about communication not spontaneous and circumstantial but rather mindful and intentional.

Explanations are a kind of social interaction and, as such, insights from psychology, sociology, and philosophy are crucial. This article highlights the relevance of a dialogue between computer science and other disciplines. Based on connecting insights from these fields, we present a *communication dynamics model* through which we can closely examine explanations within the human–AI interaction process, reflecting on the sender's explanation intention and strategy, as well as on the receiver's perception of explanation effects.

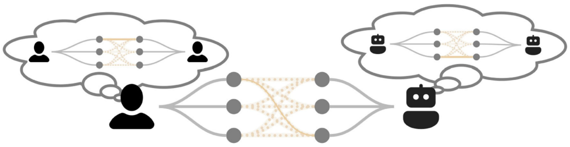


FIGURE 2. Human–AI communication. Each agent assumes a certain communication dynamic, which can lead to miscommunication if the dynamics misalign.

Based on our model, we identify and structure potential miscommunication problems between senders and receivers; in particular, we examine six reasoning pitfalls and 13 related biases. Addressing these pitfalls, we deduce six desiderata for human-centered explanations. We exemplify the implication of these pitfalls and biases in the context of an application for medical autodiagnosis. Finally, we conclude with a discussion of lessons learned, and research opportunities.

This article aims to provide a high-level viewpoint on the communication process between humans and AIs, highlighting the roles of visual and interactive explanations. Our goal is to offer a unified model with consistent terminology, which can enable researchers to provide effective design recommendations. Our model relies on analyzing best practices from the research fields of visualization, human–computer interaction, interactive and explainable AI, as well as AI philosophy.

MENTALIZING AND PERSPECTIVE-TAKING

An essential part of successful communication is mentalizing and perspective-taking. Mentalizing refers to our understanding of the inner state of minds in ourselves and others.[6] It allows us to see the points of view of other people we interact with. It also enables perspective-taking, i.e., putting ourselves in the position of others to find common grounds and knowledge gaps, enabling us to build effective argumentation and rhetorical strategies. As this process gives us the capacity to identify the knowledge gap of our counterparts, it allows us to figure out the appropriate modality and pace for communication. We explain our mental models to others in order to *teach* them about our understanding. In turn, we *learn* to change our understanding of the world and our surroundings by adapting our mental models.[21]

The processes of mentalizing and perspective-taking are part of humans' social and emotional intelligence. To

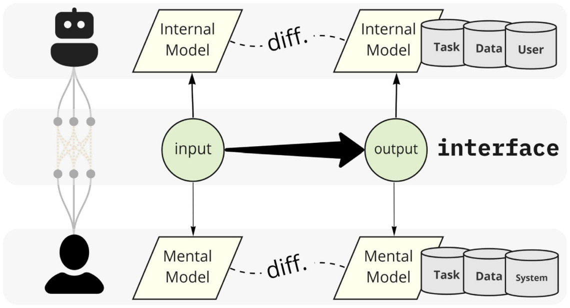


FIGURE 3. Internal and *mental model adaptation*. The model that the receiver has of the sender updates through explanations given by the latter.

analyze whether AIs can be enabled to mimic such processes, we need to further inspect the details of communication dynamics. In particular, we need to investigate *changes in mental models, triggered through explanations and interactions*.

We refer to *mental models* as the internal knowledge representations of humans. In turn, for an AI, we call the knowledge representation models *internal models*. In a mixed-initiative application context, the stored knowledge can be abstractly represented as knowledge about the *data* and the *tasks*. In addition, to adapt and evolve their understanding, users need a representation of the *AI* and AIs need a representation of the *user*. Hence, these models depict what each agent knows about its interlocutor. Figure 3 represents the adaptation of these mental and internal models through interactions and explanations on the human–AI communication interface. Through observing a transformation from an input to an output on the interface, the agents can infer information, observe patterns, and adapt their understanding. The change in their internal and mental models is the result of their communication.

These internal and mental models, however, are typically neither complete nor correct. Mental models are constantly evolving and may include nonaccurate knowledge or beliefs, acquired by agents during the learning phase. These models may provide simplified explanations to complex phenomena, used to solve problems quickly and save cognitive energy.[8] As a consequence, the inaccuracy and oversimplification of these models may lead to biases and errors in communication.[17] The model that the receiver has of the sender updates through the explanations given by the latter. To avoid miscommunication, the sender needs to be able to leverage the internal or mental model of the receiver when providing an explanation. In the next section, we explore in more detail the dynamics of communication following the process from the sender's intended internal and mental models to the receiver's perceived ones.

COMMUNICATION DYNAMICS MODEL

This section introduces our *communication dynamics model*. In the description of the model, we focus on the features of explanation as a *process* rather than a *product*. In particular, we consider how the internal/mental models affect the reception of the explanation and how the explanation, in turn, affects the updating of knowledge representations. The model describes the different communication phases between two dynamic agents: a *sender* and a *receiver*. From a high-level perspective, this dynamics can be divided in three main stages: *Internal/Mental models*, *decision-making*, and *communication channel* (see Figure 4). These stages are mirrored; the communication starts from the sender's model and ends with the update of the receiver's model on the basis of what happens in the stages in-between. Each of them can be divided into subprocesses. Both the high-level and the low-level stages are described in detail in the definition boxes shown in Figures 5–11.

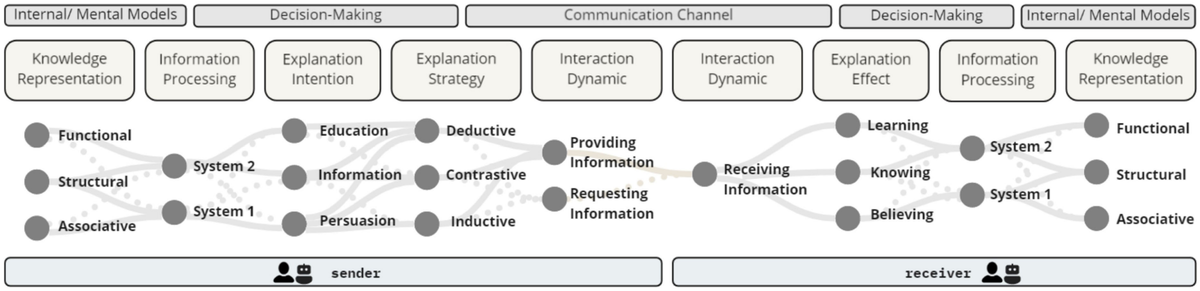


FIGURE 4. *Communication dynamics model* describing the communication stages between a sender and a receiver. Dotted lines connecting processes within the dynamics indicate a possible incorrect mapping and/or loss of information.

In the following, we describe the subcomponents of the communication dynamics model.

Sender/ Receiver

Internal/Mental Models

Internal knowledge representation about the *data* and the *tasks*. Models of the user are *mental* models, models of the AI are *internal* models.

Knowledge Representation

Decision-Making

Sender Process that starts from the analysis of the feedback received and moves through the integration of the feedback in the available knowledge base, the formation of an intention regarding the explanation that needs to be produced and ends with the choice of an explanation strategy that matches the intention.

Receiver Process that starts from acknowledging the effect that the explanation received has, and moves through the processing of the explanation, in order to include it into the knowledge representation.

Information Processing

Explanation Intention

Explanation Strategy

Explanation Effect

Communication Channel

Interaction dynamic between the sender and the receiver: while the **receiver** is always in the situation of receiving information from the sender, the **sender** may interact in order to either provide or request information.

Interaction Dynamic

FIGURE 5. Definition boxes for internal/mental models, decision-making, and communication channel.

Knowledge Representation

Expanding the mental model description from the previous section, we differentiate between the forms of information represented in the senders' and receivers' internal and mental models (see Figure 6). This includes the information that the agents have regarding the interlocutor.

Sender/ Receiver

Internal/ Mental Models

Knowledge Representation

Functional Knowing how to interact with interlocutors without knowing the mechanisms behind their behavior. It does not grant the capacity to abstract and generalize to unknown scenarios.

Structural Detailed and complete knowledge of causal and counterfactual rules about the functioning of a system. It is robust and complete and, therefore, can be used to predict behaviour in unknown scenarios.

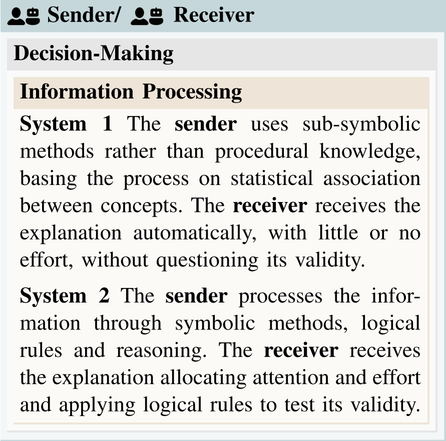
Associative Based purely on observation of statistical correlation between variables.

FIGURE 6. Knowledge representation includes the functional, structural, and associative forms of information.

The presented categorization shown in Figure 6 has been inspired by the representation of a system's competencies described by Pearl and MacKenzie in their Ladder of Causation.[18] Note that, *associative* knowledge representation can only reply to "what" questions, and factual questions, through associative reasoning, while *functional* knowledge representation can reply to "what" and "how" questions through interventionist reasoning, which considers the causes of the variable to explain. The *structural* knowledge representation is the more complete of the three. It can reply to "what," "how," and "why" questions through counterfactual reasoning. By counterfactual reasoning, we mean the capacity to reason about the causes of events in counterfactual terms (event C is said to have caused event E if, under some hypothetical counterfactual case event C did not occur, E would not have occurred).[19] It is worth pointing out here the difference that we assume between "how" and "why" explanations: while "how" explanations are not necessarily interpretable by a nonexpert audience, as they provide information about the mechanisms through which a system works, "why" explanations are the ones preferred by humans in conversational contexts.

Information Processing

To describe the mechanism of information processing, we refer to the dual-system theory of cognition, which explain the mechanism of decision making through the dichotomy between Systems 1 and 2.[5,8] System 1 processes are fast, automatic, and effortless while System 2 processes are slow, deliberate, and controlled (see Figure 7). System 1 usually offers the default and intuitive response and it is the task of System 2 to confirm or override the response by System 1. Both the sender and the receiver use these two systems when processing information for sending and receiving an explanation.


Sender/ Receiver

Decision-Making

Information Processing

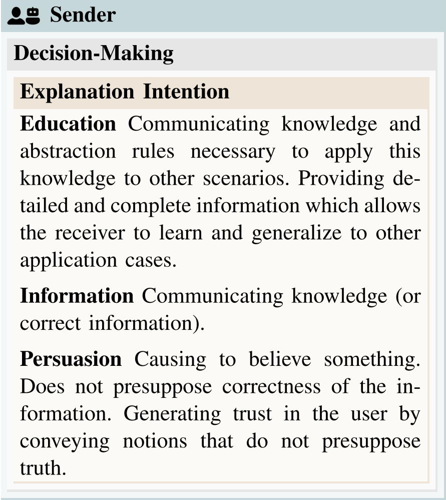
System 1 The **sender** uses sub-symbolic methods rather than procedural knowledge, basing the process on statistical association between concepts. The **receiver** receives the explanation automatically, with little or no effort, without questioning its validity.

System 2 The **sender** processes the information through symbolic methods, logical rules and reasoning. The **receiver** receives the explanation allocating attention and effort and applying logical rules to test its validity.

FIGURE 7. Information processing as explained through the dual-system theory of cognition, in which Systems 1 and 2 are contrasted.

Explanation Intention

The sender processes information and forms an explanation intention, which, in turn, affects the successive stages of the communication process (see Figure 8). Note that, when talking about intention, we do not assume any specific theory of mind, understanding it as the commitment to achieve a particular aim.


Sender

Decision-Making

Explanation Intention

Education Communicating knowledge and abstraction rules necessary to apply this knowledge to other scenarios. Providing detailed and complete information which allows the receiver to learn and generalize to other application cases.

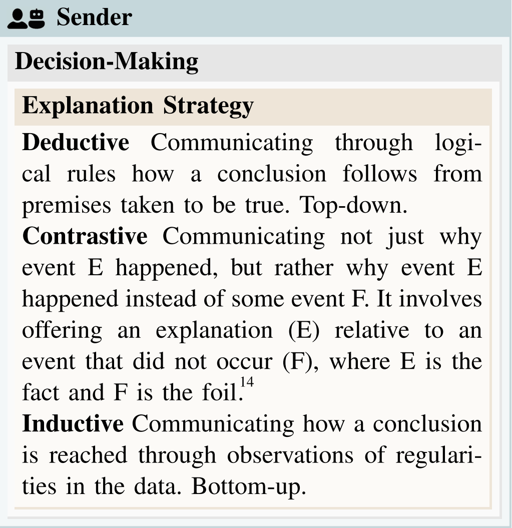
Information Communicating knowledge (or correct information).

Persuasion Causing to believe something. Does not presuppose correctness of the information. Generating trust in the user by conveying notions that do not presuppose truth.

FIGURE 8. Explanation intention of the sender, along with the successive stages of the communication process.

Explanation Strategy

Based on pedagogical research, the content of explanations can be expressed using multiple strategies and mediums[4] (see Figure 9).


Sender

Decision-Making

Explanation Strategy

Deductive Communicating through logical rules how a conclusion follows from premises taken to be true. Top-down.

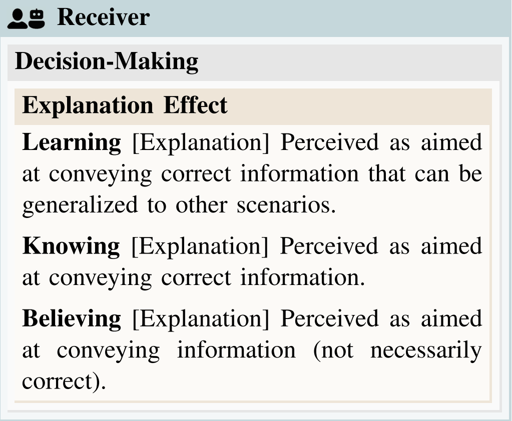
Contrastive Communicating not just why event E happened, but rather why event E happened instead of some event F. It involves offering an explanation (E) relative to an event that did not occur (F), where E is the fact and F is the foil.[14]

Inductive Communicating how a conclusion is reached through observations of regularities in the data. Bottom-up.

FIGURE 9. The explanation strategy of the sender, which can include deductive, contrastive, and inductive reasoning.

Explanation Effect

When receiving the explanation produced by the sender, the explanation causes an effect on the receiver which may, or may not, correspond to the explanation intention of the sender (see Figure 10).


Receiver

Decision-Making

Explanation Effect

Learning [Explanation] Perceived as aimed at conveying correct information that can be generalized to other scenarios.

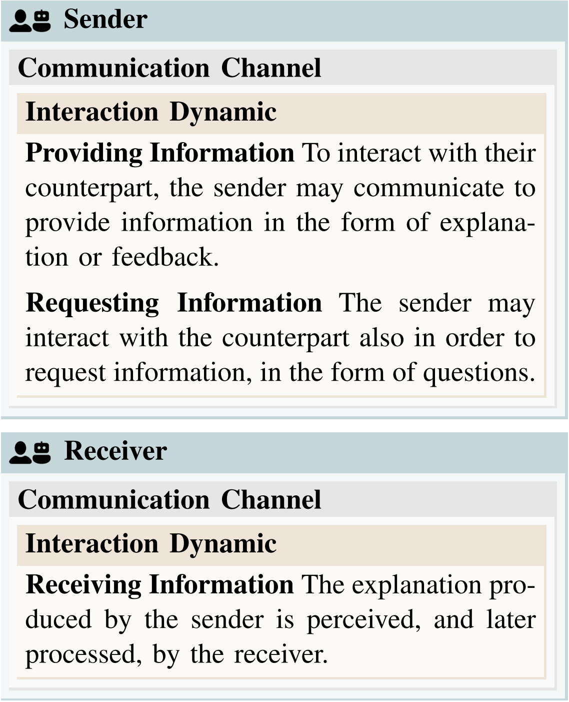
Knowing [Explanation] Perceived as aimed at conveying correct information.

Believing [Explanation] Perceived as aimed at conveying information (not necessarily correct).

FIGURE 10. The explanation effect is the characterization of the effect the sender's explanation has on the receiver.

Interaction Dynamic

The interaction dynamic is the stage of the communication dynamics where the communication between the sender and the receiver takes place and which allows the flow of information between the two agents to run (see Figure 11). The agents may use different media for actuating this process (e.g., visualization or verbalization).[4]



Sender

Communication Channel

Interaction Dynamic

Providing Information To interact with their counterpart, the sender may communicate to provide information in the form of explanation or feedback.

Requesting Information The sender may interact with the counterpart also in order to request information, in the form of questions.

Receiver

Communication Channel

Interaction Dynamic

Receiving Information The explanation produced by the sender is perceived, and later processed, by the receiver.

FIGURE 11. The Interaction Dynamic is the sender and receiver's communication dynamics as information flows between them.

BIASES AND REASONING PITFALLS

Explainees ask for an explanation mainly in two scenarios; to close a knowledge gap in their internal model, or when detecting an abnormality in a phenomenon that does not fit into the internal/mental model.

While explanations are a useful and necessary tool for learning and generalization, as they allow the receiver of the explanation to update their knowledge representation of the sender, they can also be a dangerous source of reasoning pitfalls. Inaccuracies, discrepancies, and information loss may arise in the explanation–communication dynamics when senders and receivers have different needs and intentions.

The study of how cognitive errors emerge and which social consequences they have is widely addressed in the literature in social sciences, psychology, and philosophy. Still, there is a current lack of research on cognitive biases in the context of the selection of explanations and in communication dynamics.[10] Paying closer attention to how reasoning pitfalls affect the interaction between the sender and the receiver of explanations, and the misalignment that may derive between the sender and the receiver's intentions is crucial for improving H-AI interaction.

Taxonomy of Reasoning Pitfalls

We present a taxonomy of some of the key reasoning pitfalls that can emerge in the context of a communication setting of an explanation. In the following boxes, we organize these pitfalls according to the stage of the communication dynamics in which they emerge, following the stages indicated in the scheme in Figure 4: *Internal/mental models, decision making, and communication channel*. For each pitfall, we offer some examples of related cognitive biases. We acknowledge that, by categorizing pitfalls and relative biases, we fail to capture the complexity that characterizes them and oversee some of the overlaps between the different categories.[8,15] However, this simplification is motivated by the wish to provide a consistent and accessible mapping between pitfalls and components of communication dynamics, which can be used by researchers and practitioners in debiasing techniques.

Lack of Appropriateness (Internal/Mental Models)

Inappropriate knowledge representation of sender and/or receiver, and/or inaccuracy in updating it.

Curse of Knowledge Bias: Erroneously assuming that the interlocutor has sufficient knowledge to understand.[25] What can go wrong: the sender presupposes that the receiver has background knowledge on a subject and produces an explanation that the receiver cannot understand.

Attribution Bias: Believing that the characteristics of an individual group member are reflective of the group as a whole, or vice versa. What can go wrong: the sender misinterprets the need of the receiver if it differs from the needs that are usually shared by the receiver's user group.

Availability Bias: Giving more relevance to the more easily available information. What can go wrong: the receiver updates their knowledge representation of the sender on the basis of the information about the sender that is more readily available but not necessarily the relevant one.

Lack of Interpretability (Internal/Mental Models)

The explanation cannot be processed and/or incorporated into the knowledge representation of the receiver in an accurate way.

Confirmation Bias: Accepting only the explanations that confirm previous assumptions. What can go wrong: the receiver integrates the information provided by the explanation in the set of prior beliefs without updating the knowledge representation accordingly.

Selective Attention: Attending to selected information contained in the explanation and ignoring others. What can go wrong: the receiver processes only part of the explanation received.

Lack of Informativeness (Decision-Making)

The explanation does not contribute to closing the knowledge gap of the receiver because it either provides inaccurate or insufficient information.

Recency Effect: Tendency to recall the more recent piece of information received. What can go wrong: the sender or the receiver updates their knowledge representation on the basis of the more recent feedback received by the interlocutor, disregarding past and possibly more relevant information.

Sample Bias: Selecting and presenting information in a way that fails to be representative of the phenomenon intended to be analyzed. What can go wrong: the sender provides an explanation that explains the phenomenon in a partial way.

Lack of Relevance (Decision-Making)

The explanation provides irrelevant and/or superfluous information.

Information Bias: Tendency to seek more information to improve the perceived validity of a statement even if the additional information is not relevant or helpful. What can go wrong: the receiver prefers more information to less, even if not relevant, thus perceiving the information provided by the explanation as more relevant than what it actually is.

Misinformation Effect: Tendency for postevent information to alter the original memory or knowledge of the phenomenon. What can go wrong: the receiver incorrectly processes the information without placing it in the context of previous relevant knowledge.

Lack of Accuracy (Decision-Making)

The discrepancy between the explanation intention and the explanation effect.

Authority Bias: Tendency to attribute more accuracy to the information coming from an authoritative figure. What can go wrong: the receiver who wants to fulfill the need to be educated through the requested explanation, erroneously believes that the explanation fulfills this need without questioning it if it comes from an authoritative agent.

Fair-Washing: Promoting the false perception that the explainer respects ethical values. What can go wrong: the sender manipulates the receiver's beliefs to align with their goals, with the intent of generating trust.

Lack of Interactiveness (Communication Channel)

The explanation strategy used does not support the fulfillment of the explanation intention and/or of the receiver's need.

Overconfidence Effect: Tendency of having greater confidence in one's own judgments than the objective one.[7] What can go wrong: the sender provides information through an inappropriate strategy with respect to the explanation intention, based on the confidence of the correctness of their role as an explainer.

Hyperbolic Discounting: Prioritizing immediate rewards to long-term ones, even if they are smaller. What can go wrong: the receiver accepts an explanation that does not fulfill their need in order to conclude the interaction sooner.

Contextual Evaluation of Explanations

The model of explanation–communication dynamics we propose in this article allows identifying the mechanisms that are responsible for the emergence of these reasoning and communication problems. Together with the identification of the mechanisms at the origin of these errors, a necessary first step in finding a strategy to prevent them from happening is to acknowledge the contextual and interactive nature of explanations. Different user groups vary in their needs for what should be explained and have different preferred explanation strategies.[3] An explanation for the same phenomenon may be needed by some receiver but not by others, according to whether the phenomenon is already contained in the receiver's model, and, even more basically, receivers must be aware of their lack of knowledge in order to seek an explanation. If they receive an explanation when they think they do not need it, they are not going to update their knowledge representation of the system. For this reason, attention for the development of *human-centered* strategies of explanation is crucial.[22]

The evaluation of the appropriateness of an explanation is contextual. The lack of consensus in the literature regarding the set of properties that explanations should be evaluated against should, thus, not be too much worrying: there is not such a thing as a "good"

TABLE 1. Desiderata of human-centered explanations.

Appropriateness	The explanation is appropriate to the user group, i.e., the sender has the appropriate knowledge representation of the receiver.
Interpretability	The explanation is accurately incorporated into the knowledge representation of the receiver, who updates their internal/mental model accordingly.
Informativeness	The explanation provides the necessary and sufficient information to close the receiver's knowledge gap.
Relevance	The explanation does not provide irrelevant or superfluous information that is not necessary to close the receiver's knowledge gap.
Accuracy	The effect produced by the explanation in the receiver is consistent with the explanation intention of the sender.
Interactiveness	The explanation supports effective interaction between sender and receiver, i.e., the explanation strategy is appropriate to the explanation intention and to the receiver's needs.

explanation in absolute terms, rather, there can be many good explanations, according to the target user group and their needs.

Desiderata of Human-Centered Explanations

Based on the pitfalls described previously, we deduce the following six desiderata for human-centered explanations. This list of desiderata is not intended to be exhaustive and some of the categories may partially overlap. Still, it has the benefit of offering a list of features that can help to evaluate an explanation, using the communication dynamics model we presented in Figure 4 and Table 1.

INSPECTING BIASES AND PITFALLS

The *communication dynamics model* we present in this article can be used to describe and map the mechanisms responsible for the origination of reasoning pitfalls and biases in sending and receiving an explanation. Biases are *"representative for various cognitive phenomena that materialize themselves in the form of occasionally irrational reasoning patterns, which are thought to allow humans to make fast judgments and decisions."*[10] They can involve either the sender or receiver. For example, biases may originate in the receiver, when preferring certain types of explanations over others. Explanations privilege a subset of prior beliefs, excluding the ones that are deemed inconsistent (*confirmation bias* is the most famous example: accepting explanations that confirm their assumptions). This has the advantage of reducing the cognitive load (explanations are selective) but also has the danger of perpetuating inaccuracies if explanations are generated from false beliefs.

Biases may also involve the sender, e.g., when they inaccurately update their knowledge representation due to giving more relevance to more easily available but less relevant information about the receiver (*availability bias*), or when assuming wrong features of the receiver on the basis of the data contained in the dataset (*attribution bias*).

After having presented an example of successful communication, from the biases presented in the taxonomy of pitfalls previously we select two biases: the curse of knowledge and the authority bias, showing how they can be modeled in the communication dynamics model of Figure 4 to identify the mechanisms that originated them.

Running Example: Medical Diagnosis

A dataset consisting of medical data and data of previous patients affected with COVID-19 and seasonal flu is processed through a neural network black-box method. A post-hoc XAI method based on LIME explains the decision boundary of the diagnosis generated by the black box model through an explanation based on verbalization and visualizations. The system uses natural language sentences to prompt the user to provide the necessary information for the diagnosis, for example, age, gender, past illnesses, medications, contact with infected persons, and symptoms. A chat interface enables the user 👤 to interact with the system 🤖 and to query its behavior (see Figure 12).

In this application, there can be a range of end users (e.g., AI experts or data scientists) concerned about the explainability of the model/algorithm; medical experts or physicians, concerned about clinical inference/prediction; or patients, concerned about the output reliability on the basis of symptoms and about how to proceed for curing the symptoms.

Example of Successful Interaction

Example Scenario—Starting point: The sender is the AI 🤖; the receiver is a human user 👤 The user has a functional knowledge representation: they know how to use the system, but not how it works nor the mechanism through which it produces the diagnosis. The sender still has no knowledge representation of the user.

Interaction—The user is prompted by the system in giving demographic information (age, gender, and country), medical history, and a description of the symptoms. As a result of the feedback received from the user, the system updates its knowledge representation. It has now a *structural knowledge representation* of the user, as it is able to assign the user to an end-user group and contextualize their symptoms within their demographic group, their past medical history, and the information already included in the dataset available to the system. The black-box model of the system outputs the diagnosis of "COVID." The user asks the system why it gave that particular diagnosis. The user's need is of *knowing*, of getting more details regarding how the system works. The sender then processes the information received through both Systems 1 and 2: through System 1, it correlates the user's data to similar data available in the training set to assign the user to a specific end-user group. Through System 2, it combines the information received by the user to determine the user's need. The system then forms the *explanation intention of informing* the user and it does so through a *contrastive strategy:* explaining through natural language sentences that the symptoms described correlate in 98% of cases to a diagnosis of COVID and that if the user did not have the symptoms "Loss of smell and taste" and "Chest pain," it would have produced a negative diagnosis. This information satisfies the user's need of knowing. The user processes this information through System 2, attentively considering whether the explanation is accurate and consistent with their needs, and updates the knowledge representation accordingly. The user still does not possess a full structural knowledge representation of the system, as they do not know how it works in detail, but they have a richer functional knowledge representation than what they had prior to the interaction.

Example of Sender Bias and Pitfall

We use as an example interaction between a human user and the application for autodiagnosis, described in Figure 12, to show which mechanisms of the communication dynamics are involved in the emergence of the curse of *knowledge bias*.

Example Scenario—Starting Point: The sender is the AI 🤖; the receiver is a human patient 👤 with no background knowledge about AI models and, thus, with an associative knowledge representation of the AI. The AI has no knowledge representation of the user at the beginning of the communication dynamics.

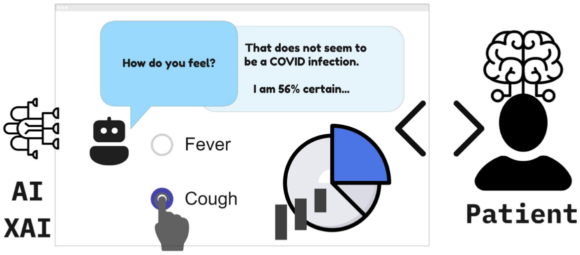


FIGURE 12. Example of a medical diagnosis system based on an interactive and explainable interface powered by both verbalization and visualization.

After the patient has provided their details and symptoms, the system 🤖 *outputs the diagnosis of "COVID-19." The patient then prompts the system, asking* why *it gave that diagnosis and, in reply, the system shows the patient a visualization of the learned deep representations of the black-box AI model, erroneously assuming that the receiver possesses the necessary background knowledge to understand it. This brings to an interruption of the communication flow, as the receiver can neither process nor incorporate in the mental model the information given by the system.*

Interpretation—The sender possesses an inappropriate representation of the receiver, as it fails to keep into account the end-user category (see Figure 13). The knowledge representation that the sender has of the receiver is at the *associative* level, formed through a statistical correlation of data included in its training set that corresponds to the profile of the patient, without taking into account the background knowledge and the end-user category of the patient. The AI also fails to integrate the feedback received by the patient through the initial screening questions and to update the knowledge representation accordingly, as it processes the new information received through System 1, i.e., automatically, without noticing the inconsistencies between the information given by the patient and the knowledge representation of the latter. By mistakenly assuming that the receiver has background knowledge in ML models, the explanation intention of the AI is to educate the receiver by showing a visualization of the features correlated with the diagnosis, an *inductive explanation strategy* that is not apt to satisfy the receiver's need. As a consequence, the flow of communication is interrupted and the patient stops using the autodiagnosis tool.

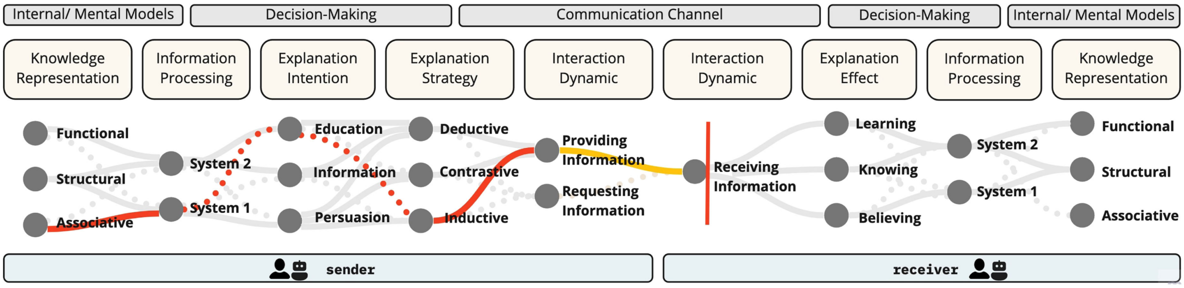


FIGURE 13. *Curse of knowledge:* bias that arises when an agent incorrectly assumes that the interlocutor has sufficient knowledge to understand. The communication flow is interrupted as the receiver cannot process the information produced by the sender.

Example of Receiver Bias and Pitfall

Using as an example another possible interaction between a user and the diagnostic AI, we describe at which stages of the communication dynamics of an explanation the *authority* bias emerges.

Example Scenario— Starting Point: Sender is the AI 🤖; receiver is a human programmer 👤 with a functional knowledge representation of the sender. The AI has a functional knowledge representation of the receiver at the beginning of the communication dynamics.

The user gives the AI their details and symptoms, and the AI 🤖 outputs the diagnosis of "Seasonal Flu." The receiver then prompts the AI asking to explain the reason behind the diagnosis, and the AI provides them with a visual explanation based on projecting the model's embedding space. What is important to note, is that the black-box model that is processing the data has a bias toward false negatives (for example, overestimating or underestimating the weight of a feature for the diagnosis) and the uncertainty-based visual encoding is not perceived by the receiver.

Interpretation—There are no inaccuracies in the communication process of the sender (see Figure 14). The sender has a correct knowledge representation of the receiver, uses the right kind of information processing with the intention of providing the receiver with information, and does so through an appropriate explanation strategy. The problems arise when the explanation is received by the programmer. Instead of processing the information received with System 2, and thus analyzing whether the diagnosis has been produced with a high level of confidence or can instead be a case of false negative, the receiver processes the information through System 1, assumes that the performance of the system is correct, and gives as feedback to the AI the recommendation of *not* suggesting to perform additional tests to patients that receive a negative diagnosis. By failing to incorporate into the knowledge representation the information regarding the bias toward false negatives, information that would allow the receiver to know how the system would behave when provided with information from other patients, the high confidence in the performance of the system makes the receiver perceive they have learned something from the system, while instead they only have a *functional knowledge representation* of it. In communication dynamics, there is a discrepancy between the explanation intention and explanation effect. In addition, the receiver

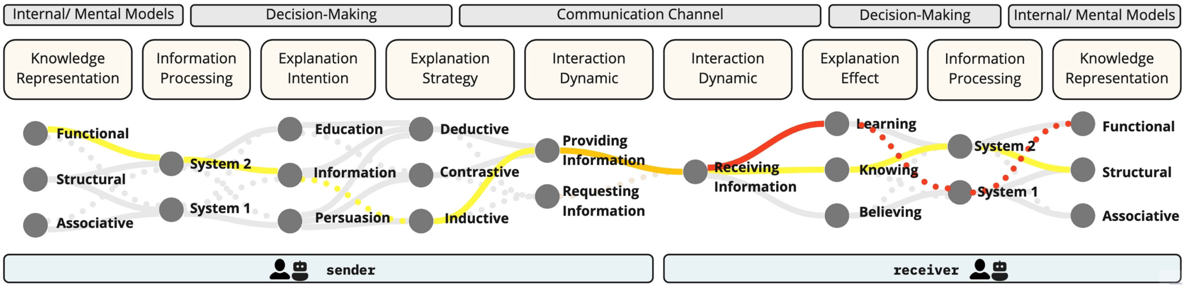


FIGURE 14. *Authority bias:* Tendency to attribute more accuracy to the opinion given by an authoritative figure. The receiver fails to update the knowledge representation accurately, assuming that the performance of the sender is correct.

inaccurately processes the explanation and, as a consequence, does not update the mental model appropriately, deeming the explanation as more complete and relevant than what it actually is.

DISCUSSION

In this article, we proposed a framework for unifying ongoing, but disconnected discourses in the literature on XAI on criteria for explanation evaluation, H-AI communication dynamics, influences of users' background on the explanation reception, taxonomies of reasoning pitfalls, and explanation kinds, by performing a decomposition of mechanisms in the communication dynamics of explanations and an analysis of which kind of errors can originate from misalignment and inconsistencies in each of the components of the explanation process. This is a necessary first step toward finding a consensus regarding the desiderata for human-centered explainable systems that can foster an effective H-AI interaction and facilitate human understanding.

The reflection on the dynamics of H-AI communication in the process of sending and receiving explanations conducted in this paper has led to various lessons learned and opened up the field to possible future research paths. In the following, we discuss three main points in each subsection.

Lessons Learned

Rational Interactions and Explanations

Humans are not always rational in their decision-making, and we do not expect them to be when we interact with them. However, an AI gives the illusion to be rational, while, in fact, it is just reflecting patterns in historical data. The *communication dynamics* model proposed provides a structure through which to map kinds of explanations to the knowledge representation through which the explanations are produced (sender) and received (receiver). Using this model for conducting an analysis of communication scenarios between humans and AI agents can help researchers to understand whether users incorporate in their mental model explanations from AI using the same conceptual framework they use for humans.

Co-Adaptation of Two Agents

The sender and receiver are two dynamic agents that interact in the process of producing and receiving explanations. The analysis of explanation as a *process* is useful to study where the errors of communication emerge and to identify areas of incorrect mapping and possible loss of information. The inspection of mechanisms of co-adaptation[20] in our model allows shedding light on whether AIs can mimic mentalizing processes, typical of successful human interactions.

Social and Emotional Intelligence

As a first step to resolving these problems, our approach aims to facilitate identifying problems in the communication between senders and receivers. Understanding errors in communication and the incorrectness that they originate in the models can help designers create systems that minimize the chances of originating these errors, thus contributing to moving toward systems that possess a higher level of social and emotional intelligence by providing a representation of dynamics of interaction in explanation processes.

Future Research Directions

Definition of Concepts

a) *Agency:* Through the *communication dynamics model* presented, it is possible to investigate the notion of agency by considering whether in an H-AI context it is always humans who assume the *initiative*, i.e., the guiding role in the conversation or whether, also, AIs can be considered to be exhibiting a certain degree of agency. Hence, assuming that being able to engage in communication dynamics and to update the knowledge representation model of the interlocutor is enough for exhibiting agency would open the possibility of considering also AIs as "agents."

b) *Understanding:* In order to allow for understanding, explanations need to be accommodated in the context of prior beliefs and not be inconsistent with those. The process of closing a knowledge gap by updating the knowledge representation and the mental model of the interlocutor can be the subject of further investigation in the search for a working definition of understanding.

c) *Explainability versus interpretability:* A possible disambiguation between the notions of explainability and interpretability may follow from the above-mentioned considerations if an interpretable explanation is understood as an explanation that allows for understanding.[9]

Empirical Research

a) *Validation through human-subject experiments:* The desiderata of human-centered explanation will enable the evaluation of the performance of XAI models.[22]

b) *Debiasing techniques:* The communication dynamics model proposed and the mapping of the reasoning pitfalls that can emerge in the process of communication can be used as a starting point to consider possible debiasing techniques that prevent these errors from happening.[10] In particular, through using effective visualization techniques, we can mitigate possible biases.[2]

Communication and Cognitive Processes

a) A more detailed study of the communication dynamics happening in the production and reception of an explanation can be conducted starting from an extended analysis of pitfalls, which includes also inductive biases and other kinds of cognitive errors.[5] This exploration can lead to a more sophisticated study of the possible overlaps between the different categories of pitfalls and desiderata identified in this article.

b) The proposed model can be used to explore further areas of research (e.g., the role of uncertainty and causal inference in communication).

CONCLUSION

To enable true *hybrid intelligence* through mixed-initiative systems, explanations and interactions are at the utmost importance. This paper presented a *communication dynamics model*, examining the impact of the sender's explanation intention and strategy on the receiver's perception of explanation effects. We provided a detailed inspection of the process of communication as an essential ingredient for successful human–AI collaboration and interaction. To that end, we also presented potential biases and reasoning pitfalls. Finally, we presented six desiderata for human-centered XAI and discussed future research opportunities.

REFERENCES

1. D. Dellermann et al., "Hybrid intelligence," *Bus. Inf. Syst. Eng.*, vol. 61, no. 5, pp. 637–643, 2019.
2. E. Dimara, S. Franconeri, C. Plaisant, A. Bezerianos, and P. Dragicevic, "A task-based taxonomy of cognitive biases for information visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 2, pp. 1413–1432, Feb. 2020.
3. U. Ehsan et al., "The who in explainable AI: How AI background shapes perceptions of AI explanations," 2021, *arXiv:2107.13509*.
4. El-Assady et al., "Towards XAI: Structuring the processes of explanations," in *Proc. ACM Workshop Hum.-Centered Mach. Learn.*, vol. 4, 2019.
5. J. S. B. T. Evans, "Dual-processing accounts of reasoning, judgment, and social cognition," *Annu. Rev. Psychol.*. vol. 59, pp. 255–278, 2008.
6. P. Fonagy, G. Gergely, E. L. Jurist, and M. Target, *Affect Regulation, Mentalization, and the Development of the Self*. London, U.K.: Routledge, 2018.
7. M. Hilbert, "Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making," *Psychol. Bull.*, vol. 138, no. 2, p. 211, 2012.
8. D. Kahneman, *Thinking, Fast and Slow*. London, U.K.: Macmillan, 2011.
9. F. C. Keil, "Explanation and understanding," *Annu. Rev. Psychol.*, vol. 57, pp. 227–254, 2006.
10. T. Kliegr, Š. Bahník, and J. Fürnkranz, "A review of possible effects of cognitive biases on interpretation of rule-based machine learning models," *Artif. Intell.*, vol. 295, 2021, Art. no. 103458.
11. T. Kulesza et al., "Tell me more? The effects of mental model soundness on personalizing an intelligent agent," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2012, pp. 1–10.
12. P. Lipton, "What good is an explanation?," in *Explanation*, Dordrecht, The Netherlands: Springer, 2001, pp. 43–59.
13. T. Lombrozo, "The structure and function of explanations," *Trends Cogn. Sci.*, vol. 10, no. 10, pp. 464–470, 2006.
14. P. Lipton, "Contrastive explanation," *Roy. Inst. Philosophy Suppl.*, vol. 27, pp. 247–266, 1990.
15. Y. Maruyama, "Rationality, cognitive bias, and artificial intelligence: A structural perspective on quantum cognitive science," in *Proc. Int. Conf. Hum.- Comput. Interact.*, 2020, pp. 172–188.
16. T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, 2019.
17. D. A. Norman, "Some observations on mental models," in *Mental Models*. London, U.K.: Psychol. Press, 2014, pp. 15–22.
18. J. Pearl and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*. New York, NY, USA: Basic Books, 2018.
19. N. J. Roese, "The functional basis of counterfactual thinking," *J. Pers. Social Psychol.*, vol. 66, no. 5, p. 805, 1994.
20. F. Sperrle et al., "Co-adaptive visual data analysis and guidance processes," *Comput. Graphics*, vol. 100, pp. 93–105, 2021.
21. F. Sperrle et al., "Learning and teaching in co-adaptive guidance for mixed-initiative visual analytics," in *Proc. EuroVis Workshop Vis. Anal.*, 2020, pp. 61–65.
22. F. Sperrle et al., "A survey of human-centered evaluations in human-centered machine learning," *Comput. Graphics Forum*, vol. 40, no. 3, pp. 543–568, 2021.
23. T. Spinner, U. Schlegel, H. Schäfer, and M. El-Assady, "explAIner: A visual analytics framework for interactive and explainable machine learning," *IEEE Trans. Vis. Comput. Graphics*, vol. 26. no. 1, pp. 1064–1074, Jan. 2020.
24. W. Willett et al., "Perception! immersion! empowerment! superpowers as inspiration for visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 1, pp. 22–32, Jan. 2022, doi: 10.1109/TVCG.2021.3114844.

25. C. Xiong, L. Van Weelden, and S. Franconeri, "The curse of knowledge in visual data communication," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 10, pp. 3051–3062, Oct. 2020.

MENNATALLAH EL-ASSADY is a research fellow at the ETH AI Center, Zürich, Switzerland. Her research focuses on combining data mining and machine learning techniques with visual analytics, specifically for text data. She is the corresponding author of this article. Contact her at menna.elassady@ai.ethz.ch.

CATERINA MORUZZI is currently a research associate with the Department of Philosophy, University of Konstanz, Konstanz, Germany. Her research interests include computational creativity, the investigation of the key guiding principles shared by human and artificial brains, and the empirical investigation of public perception of AI applications in the creative sector. Moruzzi received her Ph.D. degree in philosophy from the University of Nottingham, Nottingham, U.K. Contact her at caterina.moruzzi@uni-konstanz.de.