Chapter 11
Human-AI Interfaces are a Central Component of Trustworthy AI

Markus Plass, Michaela Kargl, Theodore Evans, Luka Brcic, Peter Regitnig, Christian Geißler, Rita Carvalho, Christoph Jansen, Norman Zerbe, Andreas Holzinger, and Heimo Müller

Abstract This chapter demonstrates the crucial role that human-AI interfaces play in conveying the trustworthiness of AI solutions to their users. Explainability is a central component of such interfaces, particularly in high-stake domains where human oversight is essential: justice, finance, security, and medicine. To successfully build and communicate trustworthiness, a user-centered approach to the design and development of AI solutions and their human interfaces is essential. In this chapter, we explain how proven methods for stakeholder analysis and user testing from human-computer interaction (HCI) research can be adapted to human-AI interaction (HAII) in support of this goal. The practical implementation of a user-centric approach is described within the context of AI applications in computational pathology.

11.1 Introduction

The prevalence of Artificial Intelligence (AI) in daily life is ever-increasing. It is integrated into smartphones and consumer goods, transforming the role of the user (Harper et al. 2020) and the human-machine interface. While the traditional human-computer interface simply represents the input-output (I/O) surface of a device (Holzinger 2004), or web page (Ebner et al. 2007), human-AI interfaces transcend the simple I/O paradigm. Besides enabling intelligent interaction via voice or facial recognition, human-AI interfaces can learn from users' behavior, react adaptively, and make predictions about future actions (Holzinger et al. 2022). Accordingly, the scope and challenges of Human–AI Interaction (HAII) research (Xu et al. 2021) differs from that of the traditional field of Human–Computer Interaction (HCI) (Dix et al. 1993). For example: AI chatbots can express human-like communication behavior (Przegalinska et al. 2019); AI-based natural language translation systems show contextual understanding (LeCun et al. 2022); AI-based programs for music co-

M. Plass (✉) · M. Kargl · T. Evans · L. Brcic · P. Regitnig · C. Geißler · R. Carvalho · C. Jansen · N. Zerbe · A. Holzinger · H. Müller
Medical University Graz, Graz, Austria
e-mail: markus.plass@medunigraz.at

M. Mehta et al. (eds.), *Explainable AI: Foundations, Methodologies and Applications*, Intelligent Systems Reference Library 232,
https://doi.org/10.1007/978-3-031-12807-3_11

creation can generate non-deterministic output (Louie et al. 2020); AI systems can collaborate with humans in teams (Calero Valdez et al. 2012; Robert et al. 2016), augment human intelligence (Crisan and Correll 2021; Holzinger 2016), and continuously learn from user behavior (Ortigosa et al. 2014).

AI has the potential to bring about a range of benefits to society, support individual and social well-being, enhance innovation and progress, and help to realize sustainable development goals (European Commission, Directorate-General for Communications Networks, Content and Technology, 2019). Regarding the case in point, AI applications in healthcare support personalized and precision medicine, drug development, critical surgeries, clinical decision and diagnosis support, medical image processing, and early detection of disease (Rajpurkar et al. 2022).

However, alongside these opportunities, the broadening application of AI brings novel risks and side effects. Fear of negative consequences, whether misplaced or valid, may also result in underuse and/or over-regulation of AI systems, leading to opportunity costs for individuals and societies (Floridi et al. 2018). Therefore, both benefits and risks must be addressed adequately to give people and societies the confidence to accept AI-based solutions, and to trust in their development, deployment, and usage, even in areas where stakes are high, such as medicine, justice, finance, and security. The trustworthiness of AI systems is a prerequisite for their uptake (European Commission 2021).

According to the *High-Level Expert Group on AI*, established by the European Commission in 2018, trustworthy AI has three components (European Commission, Directorate-General for Communications Networks, Content and Technology 2019):

(a) it should be compliant with the law
(b) it should be robust (i.e., safe, secure, and reliable to not cause unintentional harm)
(c) it should be in alignment with the four ethical principles respect for human autonomy, prevention of harm, fairness, and explicability.

This book chapter illustrates the central role that explainability and human-AI interfaces play in realizing, communicating, and verifying the trustworthiness of AI systems and the importance of a user-centered approach to the design and development of these components. The next section describes regulatory requirements for trustworthy AI and the role of human-AI interfaces in fulfilling these. Section 11.3 discusses explicability as one of the core components of trustworthy AI, and demonstrates that explainable AI is key to building trustworthiness. Section 11.4 explains why a user-centered approach is essential for achieving highly explainable and trustworthy AI systems and introduces stakeholder analysis, personas, and user-testing as valuable methods aiding user-centered design and development of AI solutions. Section 11.5 shows, with the aid of the use-case of AI applications in computational pathology, how these methods can be applied to develop human-AI interfaces that support trustworthiness.

11.2 Regulatory Requirements for Trustworthy AI

As described above, one of the three components of trustworthy AI is compliance with the law (European Commission, Directorate-General for Communications Networks, Content and Technology 2019). The *Artificial Intelligence Act* (European Commission 2021) proposed by the European Commission in 2021 is the first legal framework aimed specifically at fostering AI trustworthiness. It sets out requirements that are mandatory for all AI systems that pose significant risks to the health and safety or fundamental rights of persons (European Commission 2021). Communication to the users is a recurring feature in many of the requirements stipulated. Thus, human-AI interfaces play a central role in the fulfillment of these requirements, as illustrated in the following paragraphs:

Communicate the AI system's intended purpose and associated risks: Point 4 of article 9 'Risk management system' of the Artificial Intelligence Act specifies: "*The risk management measures ... shall be such that any residual risk ... is judged acceptable, provided that the AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse. Those residual risks shall be communicated to the user*" (European Commission 2021). To meet this requirement, human-AI interfaces must clearly communicate to users the intended purpose of an AI system as well as the residual risks associated with its usage.

Communicate the AI system's result and all information needed for its interpretation: Point 1 of article 13 'Transparency and provision of information to users' of the Artificial Intelligence Act states: "*AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately. An appropriate type and degree of transparency shall be ensured, with a view to achieving compliance with the relevant obligations of the user and the provider ....*" (European Commission 2021). To support these demands, human-AI interfaces must clearly communicate to users the AI system's output together with all information needed for the correct interpretation of this output.

Communicate instructions for use of the AI system: Point 2 of article 13 of the Artificial Intelligence Act demands: "*AI systems shall be accompanied by instructions for use ... that include concise, complete, correct and clear information that is relevant, accessible and comprehensible to users*" (European Commission 2021); and point 3 of article 13 of the Artificial Intelligence Act specifies in detail all information that shall be included in these instructions for use, such as for example "*identity and the contact details of the provider ... characteristics, capabilities, and limitations of performance of the high-risk AI system ... human oversight measures ... expected lifetime of the high-risk AI system and any necessary maintenance and care measures to ensure the proper functioning of that AI system ....*" (European Commission 2021). Human-AI interfaces can help to fulfill this requirement either by providing information on how to access the instructions for use or by conveying all information constituting instructions for use to the user directly.

Support human oversight of the AI system: Article 14 of the Artificial Intelligence Act calls for 'human oversight', and explicitly mentions the important role of the human-machine interface as a tool enabling humans to complete this task: "*AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools that they can be effectively overseen by natural persons during the period in which the AI system is in use*" (European Commission 2021). Point 4 of article 14 describes in detail all functionalities and features that human-AI interfaces must provide to support human oversight: According to point 4 of article 14, AI systems *"shall enable the individuals to whom human oversight is assigned to do the following:*

(a) *fully understand the capacities and limitations of the high-risk AI system ...;*
(b) *remain aware of the possible tendency of automatically relying or over-relying on the [AI system's] output ('automation bias'), ...*
(c) *be able to correctly interpret the AI system's output, taking into account in particular the characteristics of the system and the interpretation tools and methods available;*
(d) *be able to decide, in any particular situation, not to use the AI system or otherwise disregard, override or reverse the output of the AI system;*
(e) *be able to intervene on the operation of the AI system or interrupt the system"* (European Commission 2021).

Support the AI system's cybersecurity: Article 15 'Accuracy, robustness and cybersecurity' of the Artificial Intelligence Act requires that "*AI systems shall be designed and developed in such a way that they achieve, in the light of their intended purpose, an appropriate level of ... cybersecurity*" (European Commission 2021). Human-AI interfaces have important functions with respect to the AI system's vulnerability to cyber-attacks, for example, by enabling user authentication or by conveying security alerts to the user.

Communicate the AI system's accuracy: Article 15 'Accuracy, robustness and cybersecurity' of the Artificial Intelligence Act specifies "*AI systems shall be designed and developed in such a way that they achieve, in the light of their intended purpose, an appropriate level of accuracy .... The levels of accuracy and the relevant accuracy metrics ... shall be declared in the accompanying instructions of use*" (European Commission 2021). This means that human-AI interfaces shall always provide the user with information about the system's current accuracy so that the user can assess whether or not this level of accuracy is appropriate for the task at hand.

Support robustness of the AI system and prevent user errors Point 3 of article 15 of the Artificial Intelligence Act calls for an AI system's robustness and fault tolerance also specifically with respect to user errors: *"AI systems shall be resilient as regards errors, faults or inconsistencies that may occur within the system or the environment in which the system operates, in particular due to their interaction with natural persons or other systems"*(European Commission 2021). To fulfill this requirement, the human-AI interface on the one hand plays an important role in providing clear but

graceful feedback to the user when a user-error has occurred. On the other hand, as we know from usability research (Norman 2013) that the human-machine interface is also crucial for preventing the user from both conscious mistakes and unconscious slips during their interaction with the system.

11.3 Explicability—An Ethical Principle for Trustworthy AI

Compliance with the law is only one of the three components of trustworthy AI. Another is adherence to ethical principles and values, of which four are explicitly named by the High-Level Expert Group on AI: three traditional bioethics principles (human autonomy, prevention of harm, and fairness), which are in turn based on those described in the Charter of Fundamental Rights of the European Union (European Parliament, the Council and the Commission 2012), and a fourth: *explicability* (European Commission, Directorate-General for Communications Networks, Content and Technology 2019).

Explicability is a new ethics principle specifically relating to AI. It relates to the tendency for AI systems to act on the basis of complex internal processes that are invisible and/or unintelligible to humans (Floridi et al. 2018), rendering their decision-making processes difficult to understand, interpret, and explain (Holzinger et al. 2017). These are crucial issues for trustworthiness, validation, and acceptance of AI (Ziefle et al. 2013). According to Floridi et al. (2018), explicability recognizes the need to understand and hold to account the decision-making processes of AI.

To address the challenge of explicability, the field of *explainable AI* (XAI) research strives to provide insights into how a given AI model works and why it generates a particular result (Holzinger et al. 2018; Longo et al. 2020). There is a jumble of terms related to this concept in the XAI literature: with the terms explainability and interpretability often being used interchangeably (Zhou et al. 2021). Moreover, a variety of terms, including *transparency*, *accountability*, *intelligibility*, *understandability*, and *interpretability*, *comprehensibility* are used, sometimes interchangeably, sometimes with subtle differences in meaning that vary according to author. Other times, these terms are used without defining their specific meaning, or with one same term used for different meanings, or many different terms all referring to the same concept (Lipton 2018).

Gilpin et al. (2018) describe the concept of explainability as a combination of *interpretability* and *fidelity*, both of which are needed to achieve explainability. Here, interpretability refers to how understandable an explanation is for a human, and fidelity describes how accurately an explanation depicts the behavior of the AI model over the entire feature space. However, this often entails a trade-off between these two qualities, whereby it is difficult to simultaneously achieve both high interpretability and high fidelity: The most comprehensive explanation may not be easily interpreted by a human, and an intuitive explanation may not be sufficiently complete in its

coverage of other usage scenarios (Gilpin et al. 2018). To reach optimal explainability, it is, therefore, necessary to assess the relative importance of each of these explainability properties in a specific application context.

Miller (Miller 2019) states that we know from social sciences that usually "*people ask for 'everyday' explanations of why specific events occur, rather than explanations for general scientific phenomena*" and he argues that this holds also in the context of Artificial Intelligence (Miller 2019). To be useful, any explanation must fit the tasks and goals of the receiver of this explanation. Therefore, for an efficient and effective explanation component in an AI system, it is crucial to take into account **who** uses **which** type of AI-solution for **what** purpose, and **how** the human-AI interface is designed (Müller et al. 2022).

11.4 User-Centered Approach to Trustworthy AI

For achieving explainability, as a precondition to trustworthiness, it is critical to develop a profound and comprehensive understanding of the purpose and context of the AI application in question. This includes detailed knowledge of the stakeholders who need to understand and interpret the results provided. With respect to this deep understanding of stakeholders, the article 9 of the Artificial Intelligence Act mandates that *"due consideration shall be given to the technical knowledge, experience, education, training to be expected by the user and the environment in which the system is intended to be used"* (European Commission 2021). To this end, the following section describes methodologies for generating the rich stakeholder profiles necessary for meeting these requirements.

11.4.1 Stakeholder Analysis and Personas for AI

To achieve the aforementioned requirements for trustworthy AI, it is necessary to focus on users and use-cases throughout the conception, scoping, and implementation stages of AI application development. For traditional computer applications, such a user-centered approach (Holzinger et al. 2005) has gradually been adopted over the past four decades. There is a large set of proven tools and methodologies available for the user-/human-centered design of conventional computer systems (Vredenburg et al. 2002). However, due to the specific characteristics of AI systems, many of these existing HCI tools and methods will need to be adapted and extended to effectively support their human-centered design and development (Xu et al. 2021).

One of the existing methods successfully applied in user-centered design of conventional computer applications is that of *Personas*. This method was introduced for user-centered interaction design by Alan Cooper in 1999 (Cooper and Saffo 1999). Personas are hypothetical user archetypes that help designers and developers to empathize with the target users, to focus on the needs and goals of these users

throughout the product development process (Miaskiewicz and Kozar 2011; Nielsen 2018), and to ultimately design and develop effective, easy-to-use products (Nielsen 2019).

In traditional HCI, personas comprise the following aspects: context and environment, tasks and workflows, skills and knowledge, personal traits, goals and values, motivations and frustrations. To adapt the personas method to the context of HAII, three additional aspects describing the user's attitude to AI solutions should be taken into account (Holzinger et al. 2022):

(a) *Trust*—How much trust does the user have in the decisions/output of the AI system?
(b) *Acceptance*—Does the user accept (and follow) the decision of the AI system?
(c) *Assent*—Is the user willing to accept and use the support of the AI system?.

Furthermore, user requirements for AI applications go beyond the needs and requirements known to traditional HCI: i.e., those related to functionality, physiology, psychology, safety, usability, and user experience (Law et al. 2009). In HAII, additional needs related to explainability, decision-making authority, ethical issues, and emotion are also taken into account (Müller et al. 2021; Xu et al. 2021).

Based on the procedures for the creation of personas described in literature (Cooper and Reimann 2003; Holzinger et al. 2022; Nielsen 2019), a 5-step process can be applied to develop user personas for AI (see Fig. 11.1).

As described by Holzinger et al. (2022), a large part of the process of developing personas for AI is similar to 'traditional' (i.e., HCI-oriented) persona development. Details of this process dealing with aspects specific to personas for AI are described in the following paragraphs:

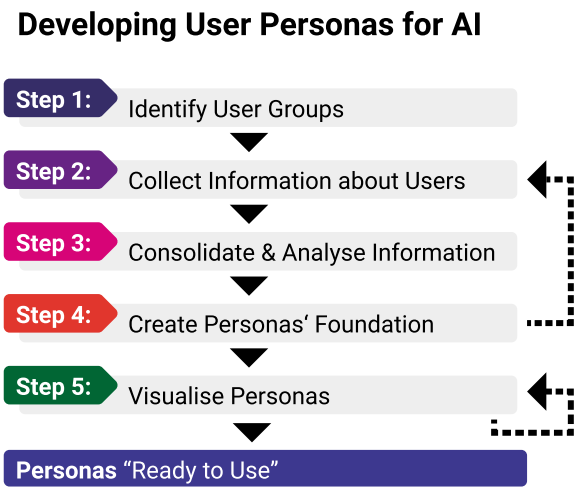

Fig. 11.1 A 5-step process, quite similar to the persona development in human-computer-interaction (HCI), can be applied to develop user personas for Artificial Intelligence (AI) based products

Step 1: Identification of (Potential) User Groups

The first step in developing personas for AI is to compile a comprehensive list of groups of people, who will potentially use the AI application. For AI applications in given a business domain, these user groups may align with job descriptions; in a consumer domain, with lifestyles (Cooper and Reimann 2003). Since each identified user group may be the seed for a distinct persona, instead of restricting the list to the most obvious end-users, a wide view should be applied in this step of the process.

For AI applications in domains where close human oversight is needed, it is necessary to include as potential users, all persons who are required to interpret and understand its results. To avoid misleading outcomes, the identification of (potential) user groups for an AI application should be data-driven. Where this is not feasible, the initial identification of (potential) user groups can be based on the assessment of domain experts.

Step 2: Collection of Information about Users This step has four distinct goals, of which the first two are also found in traditional persona development, and goals 3 and 4 are specific to personas for AI.

The first goal is to get to know (potential) users personally, discover their goals and motivations, and learn about their frustrations and hopes, their skills, education, knowledge, and personality traits. The second is to get to know the users' tasks and discover the context in which they would use the AI solution.

The third goal is specific to application cases and domains in which AI is perceived as new and innovative: find out the users' attitudes toward working with new technologies and innovations. Finally, the fourth goal is to find out the users' attitudes towards machine decisions, under which conditions they would trust a decision of an AI application, under which conditions they would follow the decision of an AI application, and whether or not they would be willing to accept support by an AI application.

Ideally, this collection of information about the users is done through ethnographic interviews or contextual inquiries (Cooper and Reimann 2003; Cooper and Saffo 1999; Pruitt and Grudin 2003). In cases, where such on-site interviews are not feasible, remote interviews should be conducted. In addition, also questionnaires or (internet) research can be utilized to complete the information.

Step 3: Consolidation and Analysis of the Collected Information

The goals of this step in the development process of personas for AI are threefold: First, to gain an overview of the collected information; second, to filter out the important findings, and third, to decide, based on these findings, which personas to develop.

The first task is to gather all collected information in one place. Depending on the kind of collected information, this central information storage can be a database or a simple spreadsheet document. It is important to take care that for each piece of information the connection to the origin is preserved throughout the whole process of organizing, structuring, splitting, and condensing the collected information.

For consolidation and analysis of the collected information various methods can be applied: visualization diagrams (such as, for example, bar charts or scatter plots) support consolidation and analysis of structured categorical or numerical information, 'affinity diagramming' helps with consolidation and analysis of unstructured information; for example, information obtained through open-ended questions in an interview or questionnaire.

These visualization and affinity diagrams demonstrate stratification within user groups, i.e., concerning features such as education, working style, personality traits, etc. It is important that each such cluster, which is related to an aspect that might influence the usage of the product, forms the basis for a persona. Clusters pertaining to user attitudes toward AI or new technologies should always be regarded as important, and should be represented accordingly in the resulting personas, as they strongly influence usage of the AI application.

Step 4: Creating the Foundation for Personas

The aim of this step in the development process of personas for AI is to create for each persona a so-called *foundation document*. This tabulates all information about a specific persona in a structured way. Various structures and templates for foundation documents of traditional personas are described in the literature (Castro and Acuña 2012; Pruitt and Grudin 2003; Pruitt and Adlin 2006). When developing a persona for AI, it is important to include in the foundation document, a specific section regarding the attitude of the persona toward AI, and toward new technology in general (potentially with notice given to whether the former category falls into the latter in the context of the application domain in question).

The purpose of the foundation document is twofold: First, the structured presentation of the collected information for a persona highlights gaps in the data where additional research may be necessary. Second, the foundation document is the basis for any usage of the persona, for example when creating a visualization of the persona or developing their use cases and scenarios.

Step 5: Visualizing Personas

The final step in this process is to transform the fictive persona into a tangible, relatable character. To bring the persona to life, it is visualized in an aesthetically appealing 1-page layout, the so-called *persona sheet*. This visualization shows the persona's name, picture, and a story conveying the persona's interests, values, lifestyle, attitudes, and behavioral patterns.

Although most of these elements are based on the information from the persona's foundation document, some fictional information may be included (e.g., regarding family or hobbies), to bring depth to the character. These fictional elements must be chosen carefully and deliberately, with the aim of supporting the communication of the persona's characteristics, whilst taking care to avoid the reinforcement of stereotypes. All important aspects of the persona described in their foundation document should be represented, in particular, regarding the persona's attitude towards AI and new technologies. Finally, to validate the visualization of a persona, it is recommended to obtain feedback from domain experts, or to show the persona sheet to

people from the respective user group and ask whether they feel plausibly and fairly represented (Marsden and Proebster 2019).

11.4.2 User-Testing for AI

As described in the previous sections, stakeholder analysis and personas are helpful methods for becoming familiar with the (potential) users and use context of AI applications, such that designers and developers may better empathize with their needs throughout the development process. However, to realize a product that is usable in practice, focusing on fictitious users is not sufficient. It is also necessary to involve real users.

User tests, e.g., Thinking Aloud methods, are proven tools and well-known from HCI and development of conventional software products (Alhadreti and Mayhew 2018; Nielsen 1993). Usually, these methods test the 'usability' of a product (Behringer et al. 2007), where this is defined by the International Organization for Standardization (ISO) as "*the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use*" (International Organization for Standardization (ISO) 2022). ISO defines 'effectiveness' as "*the accuracy and completeness with which users achieve specified goals*", 'efficiency' as "*resources used in relation to the results achieved"* and 'satisfaction' as "*freedom from discomfort and positive attitudes towards the use of the product*" (International Organization for Standardization (ISO) 2022).

However, for trustworthy AI solutions not only usability is important, but *causability* is equally crucial. Causability is defined by Holzinger et al. as *"the extent to which an explanation of a statement to a user achieves a specified level of causal understanding with effectiveness, efficiency, and satisfaction in a specified context of use"* (Holzinger et al. 2019). Thus, user tests of AI solutions should focus on not only how effectively, efficiently, and comfortably a user can achieve a specific goal using the AI solution, but also on these same criteria applied to explanations provided by the AI solution – and additionally, how *satisfied* they are with these explanations.

Aside from qualitative surveys and questionnaires (Zhou et al. 2021), causability may be quantified using the System Causability Scale (SCS) (Holzinger et al. 2020). The SCS helps determine to what extent the explanation (including process and presentation) of a result delivered by an AI solution fits the intended purpose and needs of the recipient user (Holzinger et al. 2020). When measuring causability with the SCS, the user is asked to score on a five-point scale ranging from *1=strongly agree* to *5=strongly disagree*, in response to the following ten Likert statements:

1. I found that the data included all relevant known causal factors with sufficient precision and granularity.
2. I understood the explanations within the context of my work.
3. I could change the level of detail on demand.

4. I did not need support to understand the explanations.
5. I found the explanations helped me to understand causality.
6. I was able to use the explanations with my knowledge base.
7. I did not find inconsistencies between explanations.
8. I think that most people would learn to understand the explanations very quickly.
9. I did not need more references as medical guidelines and regulations.
10. I received the explanations in a timely and efficient manner.

As with the System Usability Scale (SUS) (Lewis 2018), the final SCS Likert score is calculated as the sum of all ratings of the ten statements divided by 50 (Holzinger et al. 2020). In addition to self-reported SCS results, human mental involvement, level of understanding, emotional arousal, and stress in response to human-AI interfaces may also be measured. This analysis can be performed on eye-tracking data (Pivec et al. 2004; Preis and Müller 2003) and additional sensors such as facial expression analysis, and electrodermal activity.

11.5 An Example Use Case: Computational Pathology

Medicine is an application field in which AI solutions may bring about great benefits for individual patients, as well as public health. However, it is also a domain where stakes are high and AI solutions may introduce a high risk of harm. Therefore, as specifically mentioned in the Artificial Intelligence Act (European Commission 2021), the health sector is one of the application fields for which the trustworthiness of the implemented AI solutions is of utmost importance. Thus, we have chosen computational pathology, a subdomain of the medical imaging field, as a case study of how stakeholder analysis and the personas method may be applied to develop human-AI interfaces supporting trustworthiness.

11.5.1 AI in Computational Pathology

In histopathology, human tissue samples are investigated for cellular and/or molecular indications of diseases. In preparation, formalin-fixed-paraffin-embedded (FFPE) tissue samples are cut into ultra-thin slices, mounted on glass-slides, and pre-processed to make cellular structures and bio-markers visible under microscopy. Traditionally, these glass slides are examined by pathologists under a light microscope (Golob-Schwarzl et al. 2019; Kargl et al. 2020). In digital pathology, scanned representations of these glass slides, so-called Whole Slide Images (WSI), are examined by pathologists on a monitor (Jahn et al. 2020).

Computational pathology adds computational steps to support pathologists in their analysis of WSIs (Holzinger et al. 2017). AI for histopathological image analysis is a dynamic and growing research field (Srinidhi et al. 2021; Wulczyn et al. 2021, 2020;

Yi et al. 2019), and various AI solutions are in development to support pathologists with challenging tasks including detection of micrometastasis deposits in lymph nodes, detection and grading of prostate cancer, and immunohistochemistry-based prognostics for breast cancer (Regitnig et al. 2020).

Expectations of AI solutions in computational pathology include time savings, increasing accuracy and quality, and extraction of new medical knowledge. However, since the results of AI solutions in this area can have a tremendous impact on therapy decisions for patients, human oversight is indispensable. This renders explainability and accountability major challenges to overcome for the safe application of AI to histopathology (Holzinger et al. 2020).

11.5.2 Stakeholder Analysis for Computational Pathology

Stakeholder analysis for identifying the potential user groups of AI solutions in computational pathology is the first step in user personas development. This analysis is grounded on the question: "Who will need to understand the rationale behind the results provided by an AI solution for computational pathology and thus will need explanations for the results provided by this AI solution"?

These stakeholder groups can be identified for solutions that are in the development phase, solutions dedicated for research use only, and/or solutions approved for use in clinical work. As depicted in Fig. 11.2, these stakeholder groups include staff of the AI solution provider (software developer, quality manager, sales, customer support), staff of organizations assessing market conformity of medical software solutions (for example auditors at notified bodies designated under the EU In-Vitro Devices Regulation (IVDR) The European Parliament, The Council of the European Union 2017), staff of the pathology laboratory (pathologists, AI laboratory technician, quality manager), and researchers in medicine or molecular biology as well as researchers in data science or AI.

The need of these stakeholder groups to understand the result of an AI solution for computational pathology is based on different underlying objectives, such as debugging or improving an AI system, ensuring compliance with standards and regulations, understanding how to incorporate the AI results into further actions, and justifying or explaining actions influenced by the AI results (Suresh et al. 2021). There is therefore no one-size-fits-all solution with regards to explaining the results of AI applications in computational pathology. For example, while software developers' explanatory requirements will probably include technical details of the inner workings of the model, sales and customer support staff will usually require less technical details of the underlying algorithms but will need to understand the limits of use of the AI application and the expected accuracy of the results.

Differing levels of expected computer literacy and medical domain knowledge between stakeholder groups are illustrated in Fig. 11.3. These are important aspects to be taken into account when designing a human-AI interface or developing an explanation component for an AI solution in computational pathology.

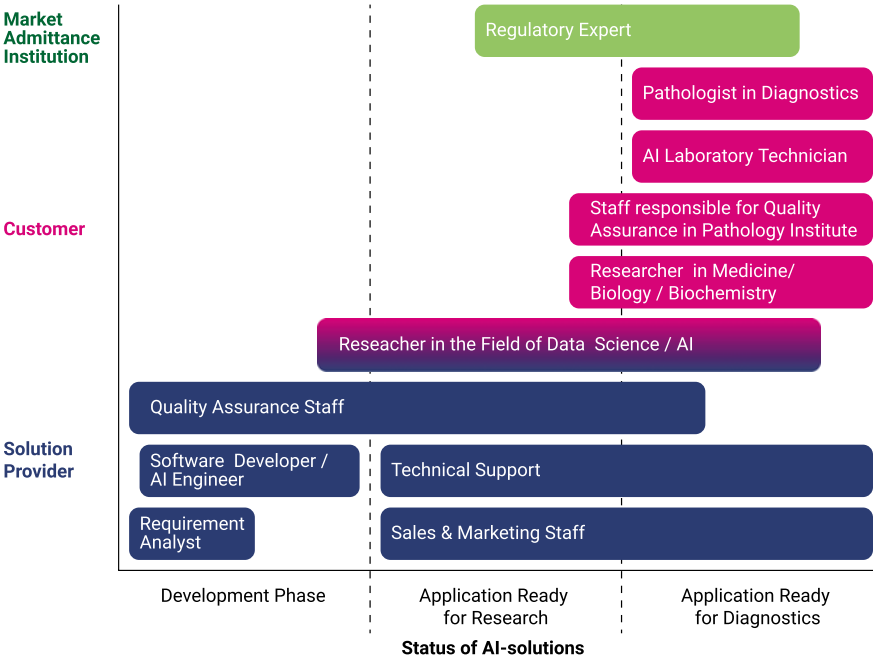

Fig. 11.2 Relevant stakeholders in different states of an AI solution for computational pathology, and their level of expertise in medicine/molecular biology

11.5.2.1 Relevant Stakeholder Groups for Computational Pathology

Groups of stakeholders who need to understand and interpret the results of AI solutions in computational pathology include staff of the solution provider, pathology institutes, market admittance institutions for medical devices, and scientists in the field of medicine and molecular biology, as well as in the field of data science and AI. All these stakeholder groups are briefly introduced in the following paragraphs.

Staff of AI Solution Providers
Roles with varying degrees of technical expertise and explainability requirements can be identified among the staff of an AI solution provider. These include requirement analysts, AI engineers and software developers, staff working in sales and marketing and technical customer support departments, as well as quality assurance managers and persons responsible for regulatory compliance.

Requirement Analysts identify the needs and demands of (potential) customers and define requirements for the to-be-developed software based on the intersection between these and the company's policies (Regitnig et al. 2020). Expertise in economics and a good understanding of the medical field in question is necessary to accomplish these tasks.

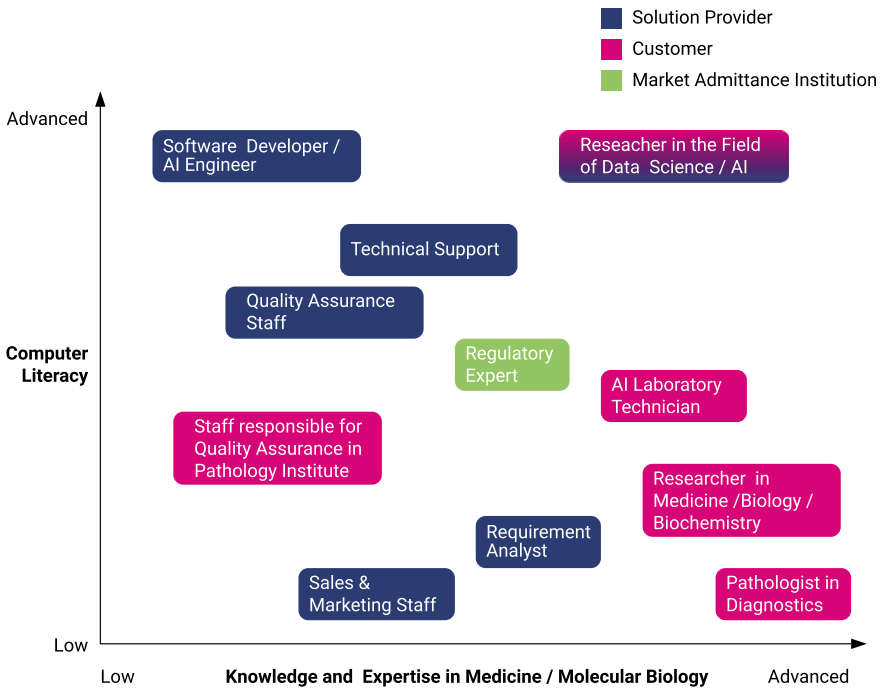

Fig. 11.3 Schematic overview of the expertise of stakeholders in computational pathology

AI Engineers and Software Developers design and implement software solutions based on the requirements collected by the requirement analyst. Typically, an education in informatics or software engineering is required, with expertise in Information Technology (IT) and computer science. They should be well aware of DICOM standards in digital pathology (Herrmann et al. 2018) and biobanking standards such as the ISO 20387 and the MIABIS ontology (Eklund et al. 2020). These stakeholders do not necessarily have extensive knowledge of medicine or molecular biology.

Sales and Marketing Staff bring the AI solution to customers. Usually, this personnel does not have extensive IT knowledge and only limited medical knowledge related to the software's application domain. However, they have a marketing and sales background and can convincingly present the solution to a (potential) customer.

Technical Support Staff are in direct contact with users and solve problems that arise during usage of the AI solution. Technical support staff often have extensive IT and/or computer science expertise, albeit without any corresponding requirement for medical domain knowledge.

Quality Assurance Manager and Person Responsible for Regulatory Compliance must have insight into the development processes and high awareness of quality standards (O'Sullivan 2019). The Person Responsible for Regulatory Compliance (PRRC), who establishes, documents, implements, and maintains a quality management system ensuring compliance with the EU In-vitro Diagnostics Regulation (IVDR), must have a degree in a relevant scientific discipline (law, medicine, pharmacy, or engineering) or four years of experience in regulatory affairs or quality management systems relating to medical devices (The European Parliament, The Council of the European Union 2017).

Staff of Pathology Institutes
Stakeholder roles with differing technical skills and needs for explainability can be identified amongst the staff of a pathology institute. The most obvious stakeholders are pathologists. However, technicians and quality managers at a pathology institute are also among those who must understand the results of AI solutions for computational pathology.

Pathologists are medical doctors who examine human tissues, cells, and body fluids in order to diagnose and monitor diseases, or predict, indicate, and monitor the outcome of therapies. Besides the findings from the microscopic examination of the specimen, a pathologist takes into account the case history and the results of other laboratory tests for these purposes. They have completed a comprehensive general medical education and (to differing degrees) highly specialized training in histopathology, and have got a strong understanding of laboratory medicine (including management, safety, and quality issues for the laboratory), excellent skills in interpreting complex patterns of test results, and knowledge regarding further tests needed for correct diagnoses.

AI Laboratory Technicians prepare AI results for pathologists. This task requires intermediate knowledge in both the IT and the medical domains. The AI laboratory technician digitizes histopathological glass slides to generate Whole Slide Images (WSIs), potentially applying pre-configured AI solutions to the resulting data. The tasks of the AI laboratory technician include the initial evaluation of AI results and adjustment of the system parameters where necessary. Furthermore, the AI laboratory technician transfers the AI results from a technical language into a format that is better suited for further analysis by pathologists.

Quality Manager at a Pathology Institute ensures the establishment, implementation, and maintenance of a quality management system compliant with standards and norms, as well as legal and regulatory requirements. The tasks of the quality manager comprise the development and monitoring of key quality indicators, key performance indicators, audit schedules, and contingency plans. Furthermore, the quality manager is responsible for risk assessment and mitigation. Typically, the quality manager at a pathology institute has a medical background and experience in quality management and economics.

Staff of Market Admittance Institutions
Every AI solution to be applied in medical diagnostics must be approved by the respective market admittance institution; i.e., the Food and Drug Administration (FDA) in the USA, or a notified body according to the Medical Device Regulation (MDR) or In-vitro Diagnostics Regulation (IVDR) in the European Union.

Auditors at a Market Admittance Institution are responsible for assessing the regulatory compliance of AI applications in medical diagnostics. Auditors at a market admittance institution usually work in interdisciplinary teams including computer experts and medical experts, such that the range of expertise necessary for the assessment of an AI solution for computational pathology is covered. While these rely predominantly on documents provided by the AI vendor to assess regulatory compliance of the product, they will also use the explanatory component of the AI solution in the evaluation of scientific validity, analytical performance, and clinical performance.

Scientific Staff
There are two groups of researchers, who may need to understand the results of computational pathology AI solutions. On the one hand, researchers in medicine or molecular biology use AI solutions to analyze human, plant, or animal specimens for their scientific work. On the other hand, there are researchers in data science or AI, who either are involved in the design and implementation of computational pathology AI solutions, or utilize (as customers) such solutions for their scientific work.

Researchers in Medicine or Molecular Biology are usually employed at a (medical) university, at the research department of a company (for example, in the pharmaceutical industry), or at a public institution (for example, the World Health organization). Typically, these researchers are not trained IT experts, but rather possess a university degree in medicine, biology, biochemistry, or pharmacy. They need AI solutions that are easily available and affordable, whilst remaining adaptable and extendable to their specific research question.

Researchers in Data Science or AI are involved in the development of AI solutions for computational pathology, trained as they are for translating real-world problems into machine learning approaches. They are skilled in designing, configuring, and adjusting complete AI solutions for discovering patterns in data. Usually, researchers in data science or AI have got an education in mathematics, computer science, or software engineering.

Aside from, or in addition to, being involved in AI development, they may use AI solutions as tools for their scientific work, and therefore may also be found on the consumer side of the AI solution life cycle.

11.5.2.2 Personas for Computational Pathology

As described in Sect. 11.4, user personas are a proven method to help understand (future) users, user interaction, and the context of use of a product, therefore supporting designers and developers in focusing on the needs and goals of their products' users.

We have developed user personas for all stakeholders of AI in computational pathology - a detailed description of this process can be found in (Holzinger et al. 2022). In the following paragraphs, the personas developed for the user group of pathologists are provided as an example.

In step 1 of the 5-step process for persona development: *Identification of (Potential) User Groups*, pathologists were identified as one of the user groups who must understand and interpret results delivered by AI solutions in computational pathology.

In step 2: *Collection of Information about the Users*, personal and work-related information was collected via contextual interviews with pathologists, an online survey among pathologists, and internet research. The collected work-related information included tasks, workflows, work context, education, experience, skills, and knowledge. Personal information comprised of goals, motivations, frustrations, personal traits, values, learning style, as well as attitudes about new technologies and AI.

In step 3: *Analysis of the Collected Information*, we recognized clusters in the answers of the pathologists with regards to two aspects that may influence pathologists' usage of AI solutions in computational pathology:

(a) some pathologists work only in diagnostics, while others (particularly in research hospitals connected to a medical university) work also in research, and
(b) some pathologists like to work with new technologies and are open-minded towards the usage of AI, while others are no technophiles and not so fond of working with new technologies.

Both of these aspects are important for the design and development of AI solutions for computational pathology: The finding (a) that some pathologists work only in diagnostics while others work also in research is relevant with respect to how an AI solution would be used by these two groups, since research work is different from routine diagnostics:

First, research work is usually multi-disciplinary and typically includes experts from various research fields such as medicine, pharmacy, biochemistry, biology, and bioinformatics. Second, research work introduces new methods and approaches and can be more exploratory and experimental than routine diagnostics, as it aims to go beyond the state-of-the-art and generate new scientific findings. Third, research work is usually conducted under less strict time constraints than routine diagnostics, therefore allowing more resources to be expended on a single medical case.

The finding (b) that some pathologists are technophiles and some are not is relevant with respect to their requirements for AI solutions: Users who are technophiles and well versed in exploring and trying out new technologies, are likely to be open to

applications that offer more options for customization, while users who are not so fond of spending time getting familiar with new technologies may require more guidance and rely on applications that offer easy-to-use default procedures.

For step 4, we created foundation documents for 3 different 'pathologist' personas based on these findings, which are shown in Table 11.1.

Finally, in the last step of the 5-step process for persona development, each of these 3 'pathologist' personas was visualized in a persona-sheet, rendering these fictional characters in a tangible and realistic way. These persona-sheets can be found in the appendix (Sect. 11.7).

11.5.3 Human-AI Interface in Computational Pathology

As so far demonstrated, it is essential to consider the users and the usage context when designing and developing human-AI interfaces. Thus, in defining the needed functionalities of a user interface (UI) for computational pathology AI solutions that aim to support pathologists with WSI analysis, it is important to take into account the workflows in pathology (Kargl et al. 2020) and the process of how pathologists develop a diagnosis (Pohn et al. 2019a, b).

With regards to functionality, the UI in computational pathology can be split into two components: (1) The basic digital pathology UI, enabling the user to manage and view WSIs, and (2) the AI-specific component, which is a human-AI interface that enables the user to interact with a specific computational pathology AI solution.

This AI-specific component can be further split into two components, based on functionality: (2a) the primary AI interface, which enables the user to request, view, and interact with the primary result of the AI solution, and (2b) the explanation component, which enables the user to request and view an explanation as required for the primary result of the AI solution, as depicted in Fig. 11.4. Ideally, these components should be integrated, with the basic UI providing the possibility to dynamically add (plugin) user interaction functionalities for specific application contexts.

11.5.3.1 Basic UI Components: Case and Slide Viewer

The *Case Viewer* and the *Slide Viewer* are at the core of the UI in computational pathology. They can be implemented as combined software or as stand-alone products.

The Case Viewer enables the user to access and view a set of WSIs, as shown in Fig. 11.5. In diagnostics, where a single medical case can comprise up to 100 WSIs, the task of the case viewer is to present all these WSIs of a medical case to the pathologist in a structured way. In research, the case viewer helps to manage and organize WSIs for projects. Frequently, the case viewer is a pathologist's entry point to digital pathology, as it forms the link between the slide viewer and the Laboratory Information System (LIMS) in a pathology institute.

Table 11.1 Foundation documents for 'pathologist' personas

	Persona 1 Pathologist in diagnostics type 1	Persona 2 Pathologist in diagnostics type 2	Persona 3 Pathologist in research
Tasks	Diagnostics macroscopic examination of specimens microscopic examination of specimens examination of frozen sections participate in tumor-boards autopsy administration	Diagnostics macroscopic examination of specimens microscopic examination of specimens examination of frozen sections participate in tumor-boards autopsy administration	Examination of WSIs for research scientific publications participate in scientific projects applying for research funding administration
Personal traits	Likes working accurately and precisely obsessed with details conscientious curios, spirit of research not very spontaneous not technophile rather reluctant, resistant to changes	Likes working accurately and precisely obsessed with details conscientious curios, spirit of research not very spontaneous likes working with new technologies open-minded to changes	Likes working accurately and precisely obsessed with detail curios, spirit of research likes working with new technologies open-minded to changes
Motivations	Enjoys pathology work finding the right diagnosis for patients solving challenging and difficult cases recognition of my work among peers	Enjoys pathology work finding the right diagnosis for patients solving challenging and difficult cases recognition of my work among peers bring new scientific findings to practice	Enjoys pathology work new scientific insights recognition of my work in scientific community
Frustrations	Feuding among colleagues unnecessary delays and waiting times administrative stuff inaccuracy and imprecision insufficient clinical information too much work disturbances at work (phone calls, people dropping in, training sessions, slow IT services, computer error...) lack of recognition of pathologists' work by the general public and by clinicians	Feuding among colleagues unnecessary delays and waiting times administrative stuff inaccuracy and imprecision insufficient clinical information too much work to be pressed for time no time for reading specialist literature	Feuding among colleagues unnecessary delays and waiting times administrative hurdles too much work lack of finance
Ideal working environment	Just let me work—no disturbances minimise administrative work amount of work just right flexible working hours, home office calm surroundings (no other people in the room) reliable infrastructure ergonomic workplace	Minimise administrative work being innovative support of employer for my visions good colleagues relaxed atmosphere at the workspace amount of work just right up-to-date infrastructure and tools flexible working hours, home office ergonomic workplace	Just let me work—no disturbances minimise administrative work being innovative support of employer for my visions good colleagues relaxed atmosphere at the workspace amount of work just right up-to-date infrastructure and tools flexible working hours, home office

The case viewer provides functionality related to the organization of cases and WSIs (e.g., grouping of WSIs in cases and blocks, reordering of WSIs, prioritization of cases, and user access management to cases), and functionality related to the enrichment of WSIs with metadata (e.g., gathering case-relevant metadata from LIMS, and triggering AI solutions to get specific WSI analysis results). Furthermore,

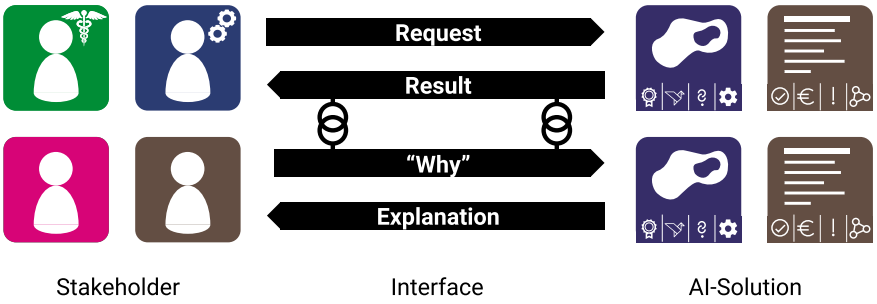

Fig. 11.4 The human-AI interface in computational pathology

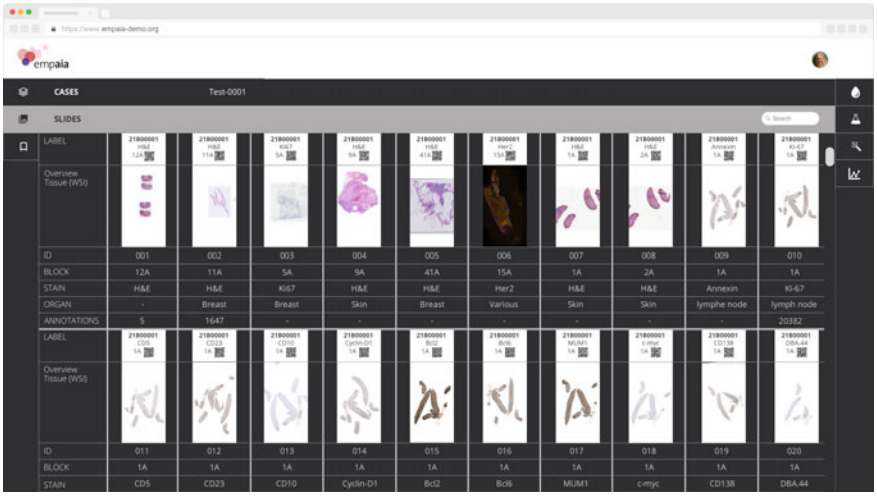

Fig. 11.5 Example of a Case Viewer, which enables the user to access, view, organize, and manage a set of whole slide images (WSIs)

the case viewer offers plugin integration, which can be used to add for, example, WSI pre-processing functionality to the case viewer.

The Slide Viewer enables the user to view a WSI and related annotations on the screen, and provides appropriate interaction techniques for such gigapixel images, as shown in Fig. 11.6. Functionalities provided by the slide viewer include: visualizing the WSI at different magnifications, creating user annotations (e.g., text annotations or marking a region of interest (ROI) with a circle, rectangle or polygon), visualizing annotations created by the user or by an AI solution, scrolling through different focus-layers of a WSI (so-called z-stacking), automated alignment of WSIs, performing color-correction and channel-adjustment for WSI, as well as tracking the user's viewing activities for a WSI and marking already-viewed areas. Furthermore, the slide viewer offers plugin integration, which can be used to add third-party applications and extend the slide viewer's functionality, for example, with a quantification tool.

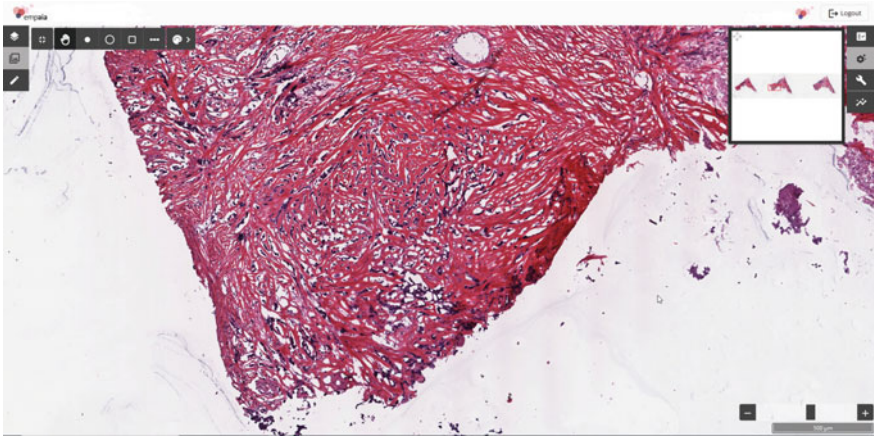

Fig. 11.6 Example of a Slide Viewer, which enables the user to view a whole slide image (WSIs) on the screen, navigate through the WSI and make annotations

11.5.3.2 AI-Specific UI Components

The AI-specific UI component comprises two functional parts: the primary AI interface providing the interaction and visualization functionalities, supporting the primary/main purpose of the AI solution, and the explanation component supporting the user's 'Why' questions (Miller 2019). Figure 11.7 shows an example of such an AI-specific UI in computational pathology.

The Primary AI Interface must, on the one hand, support communication of a specific user request to the AI solution, and, on the other hand, it must support communication of the AI solution's result to the user. To enable the user to formulate a specific request for the AI solution, the interface shall, for example, provide annotation functionality enabling the user to easily define ROIs and select specific parts of the WSI as input parameters for the AI solution.

Since pathologists are accustomed to looking at images and trained in interpreting image information, the results provided by an AI solution in computational pathology should, whenever possible, be communicated to the pathologists visually, as an overlay on the WSI, for example as pixel-based or ROI-based annotations. Ideally, different AI solutions used in a domain such as computational pathology, should apply a consistent visual vocabulary of symbols and color codes when visualizing results and/or explanations.

The Explanation Component should convey explanatory information to the user. A large variety of explanation approaches for AI exists, and a comprehensive overview of those state-of-the-art XAI methods, which are applicable to computational pathology, is given by Pocevičiūtė (Pocevičiūtė et al. 2020). Possible modalities include visual overlays on the WSI (such as shapes, heatmaps, or saliency maps), figures and

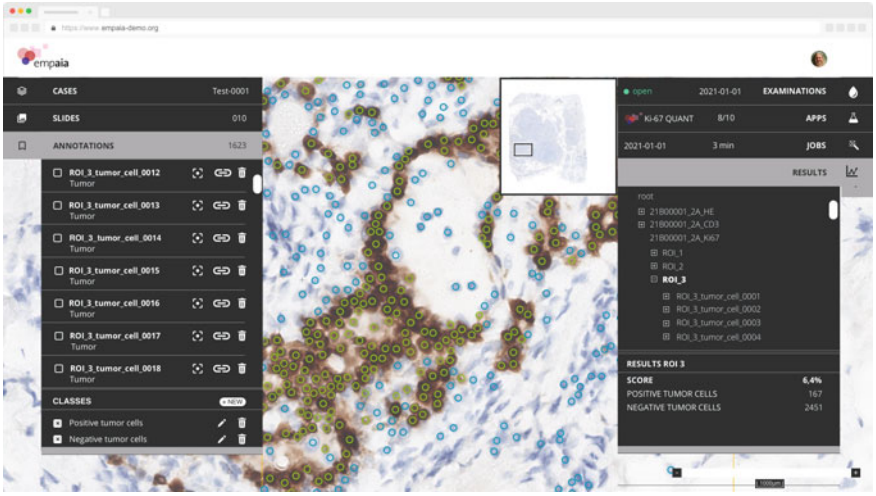

Fig. 11.7 Example of a UI displaying results and explanatory annotations generated by an AI solution. Here, the result is the overall positivity score (bottom right), whereas an intermediate result, consisting of individual cell annotations, is included as an explanatory element that helps the user to understand this outcome

charts, text labels, example images (e.g., counterfactuals or prototypes), interactive dialogue systems, or even simply intermediate results of the AI solution.

When choosing an explanation method, it must be considered that a good explanation should support the context, task, and goal of the recipient of the explanation. Thus, to determine which kind of explanation is appropriate and how the explanatory information should be provided, it is necessary to analyze the specific task and context in which a specific user requires an explanation. As many different stakeholder groups have been identified to be relevant for AI solutions in computational pathology, it should be considered to implement explanation components optimized for different user groups. When the same explanation method is used for different stakeholder groups, it is necessary to thoroughly adapt it to the knowledge and needs of the respective members. For example, since a pathologist's common task of examining WSIs to derive a diagnosis is a visual task, pathologists prefer also visual modalities for explanations of AI solution's results (Evans et al. 2022). In addition, also the simplicity of the explanation method is an important criterion for pathologists, as they usually face a high workload and a tight schedule and thus consider time savings as the most valuable benefit of AI assistance (Evans et al. 2022).

It is important that the explanation considers the user's background knowledge and uses concepts, which are familiar to the user. Moreover, an explanation must also match the user's need for information. For example, if a pathologist needs to verify the correctness of a cell ratio calculated by an AI solution, an explanation highlighting the cells recognized by the AI solution and the respective ROI may be easily understandable, since it uses concepts from the medical domain (cells) and from common knowledge (how to calculate a ratio) that are familiar to the pathologist.

However, an explanation approach displaying a saliency map of the most relevant parts of the image for a given AI output, may not help the pathologist to assess the correctness of the ratio calculated by the AI solution, and may therefore be inappropriate in this context. Such a saliency map could, however, be a helpful explanation for a software developer searching for clues to a 'Clever-Hans' problem (Pfungst 1911) in the AI model.

The explanation component of the human-AI interface is a crucial element for the trustworthiness of AI solutions in computational pathology. However, as shown by Evans et al. (2022), ambiguous explanations may also pose a significant risk of introducing inappropriate trust in AI solutions. This risk is particularly pronounced when the explanation seems to imply a causal decision-making process of the AI model similar to the user's own. For example, saliency maps pointing to diagnostically relevant regions, whilst obscuring or omitting the important features represented within these regions, or synthetically created counterfactuals with several features changing simultaneously leaving it unclear which of these were truly relevant (Evans et al. 2022).

Therefore, especially in high-risk domains such as medicine, it is important that human-AI interfaces are carefully designed and thoroughly tested with users from the target group before they are applied in practice.

11.6 Conclusion

In summary, both the design and the evaluation of the human-AI interface play a central and increasingly important role in achieving and verifying the trustworthiness of an AI solution. This is particularly relevant in application domains that directly impact human life and livelihood. In reviewing ethical criteria for AI applications in biomedical research and biobanking (Kargl et al. 2022), there is a clear need for further basic research that will then lead to concrete practical guidelines.

In the context of legislation and regulatory practice, e.g., the European In-Vitro Diagnostics Device Regulation (IVDR) / Medical Device Regulation (MDR) and U.S. Food and Drug Administration (FDA) activities, aspects of explainability and causability of the human-AI interface will be an indispensable component for the validation of critical AI solutions (Müller et al. 2022). In order to meet the stringent requirements of both extant and upcoming regulatory frameworks, further fundamental research is imperative.

Robust and explainable human-AI interfaces will be the central component for building truth and trustworthiness in AI systems, and therefore constitute an important and exciting area of future research. In particular, further work is required to make the explainability and causability of Human-AI interfaces in critical domains measurable and quantifiable. To this effect, it is necessary to develop strategies for the collection of ground truth relating to these metrics. Future dialog systems and benchmark datasets such as the Kandinsky Patterns (Müller and Holzinger 2021) can provide valuable assistance to the international research community toward this goal.

11.7 List of Abbreviations

AI	Artificial Intelligence
DICOM	Digital Imaging and Communications in Medicine (standard)
EC	European Commission
EU	European Union
FDA	United States Food and Drug Administration
FFPE	Formalin-Fixed Paraffin-Embedded
HAII	Human-AI Interaction
HCI	Human-Computer Interaction
ISO	International Organization for Standardization
IT	Information Technology
IVDR	European In-Vitro Devices Regulation
I/O	Input-Output
LIMS	Laboratory Information System
MDR	European Medical Devices Regulation
MIABIS	Minimum Information About Biobank data Sharing (standard)
PRRC	Person Responsible for Regulatory Compliance
ROI	Region of Interest
SCS	System Causability Scale
SUS	System Usability Scale
UI	User Interface
WSI	Whole Slide Image
XAI	Explainable Artificial Intelligence

Acknowledgements Parts of this work have received funding from the European Union's Horizon 2020 research and innovation program under grant agreements No. 857122 (CY-Biobank), No. 824087 (EOSC-Life), and No. 874662 (HEAP). This publication reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains. Parts of this work have received funding from the Austrian Research Promotion Agency (FFG) under grant agreement No. 879881 (EMPAIA) and by the Austrian Science Fund (FWF), Project: P-32554 explainable Artificial Intelligence. Part of this work was done in the context of EMPAIA, a project funded by the German Federal Ministry for Economic Affairs and Climate Action, with funding codes 01MK20002A, 01MK20002C, and 01MK20002E. We are very grateful for the proofreading of the manuscript by Bettina Kipperer.

Appendix

Three sample personas for AI applications in digital pathology, as described in Sect. 11.5.2.2 (Figs. 11.8, 11.9 and 11.10).