Case 1: NarxCare algorithms

Q1 (Transparency): Which approaches towards transparency do you think would be needed to ameliorate unjust outcomes ensuing from the use of NarxCare algorithms (see Kathryn's case)?

To ameliorate unjust outcomes from the use of the NarxCare algorithm, the following approaches to **transparency** should be adopted:

1. **Value Transparency**: Clearly communicate the objectives and values the algorithm is designed to optimize. This would allow stakeholders, such as physicians and patients, to understand the motivations and trade-offs that underlie the algorithm's decisions【75†source】.
2. **Translation Transparency**: Provide information on how the real-world inputs (e.g., a patient's medical history) are translated into features that the algorithm uses. This could help physicians understand how data points are processed and how specific inputs influence the risk score【75†source】.
3. **Performance Transparency**: Communicate the accuracy, false positive rates, and any limitations of the algorithm in predicting addiction risk. This would allow physicians to better judge when to trust the output of the algorithm and when to question it【75†source】.
4. **Consistency Transparency**: Demonstrate that the algorithm provides consistent outputs given similar inputs, helping to alleviate concerns about arbitrary or unfair outcomes【75†source】.

These transparency measures would ensure that physicians and patients can better understand how decisions are made, reducing the likelihood of unjust outcomes due to algorithmic black-box behaviors.

Q2 (Responsibility): In your opinion, to what extent are physicians to be considered morally responsible for the decisions taken in the case under scrutiny? Can you recognize any responsibility gaps?

Physicians can be considered **partially morally responsible** for the decisions taken in Kathryn's case, but their responsibility is limited by the influence of the NarxCare algorithm. Physicians are responsible for their role in acting on the algorithm's output, as they have the professional duty to consider all available information when making a medical decision. However, they are **not solely responsible** because they may have been pressured by institutional policies to follow the algorithm's recommendations strictly, which limits their autonomy【74†source】.

A **responsibility gap** exists here because the NarxCare algorithm itself cannot be held accountable, and yet its output significantly impacts medical decisions. The algorithm developers, healthcare administrators, and policymakers share responsibility in ensuring that the tools provided to physicians are accurate, reliable, and accompanied by guidance that supports ethical use. The **lack of explainability** and the **pressure to comply** with algorithmic recommendations create gaps in accountability, as physicians cannot always critically assess the algorithm's validity【74†source】.

Q3 (Contestability and Value Conflicts): Which forms of contestability do you think should be designed into these systems to empower patients that are attributed a risk score that does not mirror their actual drug consumption levels? Can any values conflicts emerge from allowing these forms of contestability?

To empower patients who are assigned an inaccurate risk score, the following forms of **contestability** should be designed into the NarxCare system:

1. **Automated Contestability**: Allow patients to request a review of the risk score through an automated system that assesses whether relevant information may have been misinterpreted or missed【72†source】.
2. **Human Oversight and Review**: Implement mechanisms for **manual review** by human experts, where patients can appeal the risk score and provide additional context or evidence that the algorithm may not have considered.
3. **Transparent Feedback Loop**: Ensure that patients can view the **inputs** that contributed to their risk score and provide corrections if there are inaccuracies. This makes the process more transparent and allows patients to actively participate in the decision-making process.

Value conflicts could emerge from allowing these forms of contestability. On one hand, contestability promotes **fairness** and **autonomy** by enabling patients to correct potentially harmful misjudgments. On the other hand, healthcare institutions may argue that increased contestability reduces **efficiency** and can **undermine trust** in the system, especially if patients contest decisions too frequently without just cause. Balancing fairness, patient empowerment, and the efficient functioning of healthcare systems is a significant challenge【79†source】.

Q4 (Explainability): Which kind of explanatory information would be needed, in your opinion, for physicians to understand and evaluate the suitability of the outputs produced by these systems?

Physicians need the following types of **explanatory information** to understand and evaluate the suitability of the outputs produced by NarxCare:

1. **Saliency Information**: Highlight the most influential inputs contributing to a given risk score (e.g., recent prescriptions, patient history). This would allow physicians to see the factors that played the most critical role in generating the score, enabling them to make informed judgments about its appropriateness【77†source】.
2. **Decision Path Transparency**: Provide a simplified explanation of the **decision path** or **logic** that led to the given score, such as which thresholds or rules were applied. This would help physicians understand how specific factors were weighed and combined in generating the final score【77†source】.
3. **Alternative Scenarios**: Offer **what-if analysis**, allowing physicians to adjust certain inputs and observe the effect on the risk score. This would help them evaluate the sensitivity of the score to different variables, giving them insights into how robust or fragile the decision is in light of new information【77†source】.
4. **Performance Metrics**: Display metrics such as **accuracy, false positive/negative rates**, and known limitations, allowing physicians to assess the reliability of the output and make decisions accordingly. This would help them decide when it may be appropriate to override the algorithm's recommendation【75†source】.

Case 2: AI and COVID

Q1 (Transparency and Explainability): Do you think that the failures in the clinical implementation of AI systems to, for example, triage patients are attributable to limitations in

terms of transparency and explainability of these systems? If so, to what extent and why?

Yes, the failures in the clinical implementation of AI systems for triage are significantly attributable to **limitations in transparency and explainability**. These limitations hindered effective integration into clinical practice for several reasons:

1. **Lack of Algorithm Transparency**: The AI systems used were often **black-box models**, meaning that clinicians could not understand how the models arrived at their decisions. This lack of transparency made it difficult for medical professionals to trust the outputs, and prevented them from intervening when the algorithm made an error【75†source】.
2. **Explainability and Trust**: Clinicians were not provided with explanations of how decisions were made, which is critical in a high-stakes environment like healthcare. Without explainability, doctors could not validate the decisions against their clinical expertise, making them hesitant to rely on the AI outputs. The failure to generate trust among healthcare professionals was a major reason these systems were ineffective in practice【77†source】.
3. **Opaque Data and Methodology**: The quality of the training data was also opaque—data that may have been **mislabelled** or from unknown sources contributed to unreliable model predictions. When data sources are not transparent, it becomes difficult to determine the validity of the model's predictions in a real-world context【74†source】.

In summary, the lack of both **transparency** in the data and algorithmic processes, as well as the absence of **explainability** mechanisms, played a crucial role in the failures observed. This opacity ultimately led to mistrust among clinicians and a lack of proper oversight, which hindered the clinical adoption of these tools.

Q2 (Value Conflicts): Consider possible value conflicts emerging from the use of systems triaging patients. Briefly mention the nature of the conflict emerging between which values and discuss a possible trade-off strategy suitable for the case in question.

The use of systems to triage patients gives rise to **value conflicts** between **efficiency** and **equity**:

- **Efficiency vs. Fairness**: AI systems are designed to triage patients efficiently, aiming to allocate resources as quickly as possible. However, this efficiency may come at the cost of **fairness**. Vulnerable groups, such as minorities or those with atypical medical histories, may be underrepresented in training data, leading to biased triage outcomes. The conflict arises because the pursuit of efficiency can exacerbate disparities in healthcare outcomes【79†source】.
- **Trade-off Strategy**: A suitable trade-off strategy would be to implement **fairness-aware algorithms** that include specific measures to ensure that minority or vulnerable populations are adequately considered in the decision-making process. This may mean allowing for a slight decrease in triage speed (efficiency) to ensure that no group is systematically disadvantaged by the system. In this context, **satisficing** approaches, where an acceptable balance between fairness and efficiency is sought, could be implemented to resolve this conflict【79†source】.

Q3 (Responsibility): To what extent do you think that AI developers can be considered responsible for the failed implementations of AI systems in the clinical setting? Why?

AI developers bear a significant portion of the responsibility for the failed implementation of these AI systems in the clinical setting, for the following reasons:

1. **Quality of Training Data**: Developers have a responsibility to ensure the use of **high-quality, representative datasets** when training AI models. In the case described, developers used data that was either **mislabelled** or from **unknown sources**, which directly affected the reliability of the model's output【74†source】.
2. **Lack of Robust Testing**: Developers did not adequately test the systems under real-world conditions, resulting in models that were **not suitable** for clinical settings. Proper validation and robustness checks are essential before deploying an AI system in a high-stakes environment like healthcare【77†source】.
3. **Transparency and Explainability**: AI developers failed to design systems with sufficient transparency and explainability. This lack of consideration left clinicians unable to understand or challenge the outputs, effectively removing their ability to exercise their professional judgment in cases where the AI might be wrong【77†source】.

Therefore, AI developers are responsible to a considerable extent for these failures because the problems stemmed largely from shortcomings in development practices, data quality, and a lack of attention to user requirements.

Q4 (Contestability): Do you think that the possibility to contest the decisions of these systems is of paramount importance in clinical settings or are there other requirements that should be given higher priority? Please, motivate your answer.

Contestability is of paramount importance in clinical settings, but it must be combined with other requirements such as **explainability** and **data quality**:

1. **Importance of Contestability**: Given the high stakes in healthcare, the ability for healthcare professionals to **contest AI decisions** is crucial. Contestability ensures that erroneous or biased decisions can be reviewed, and patients can receive a fair assessment based on human oversight. Without this mechanism, the risk of harm due to incorrect or biased outputs is significantly increased【72†source】.
2. **Explainability as a Foundation**: However, for contestability to be effective, **explainability** must also be prioritized. Clinicians need to understand why the system made a particular decision to challenge it meaningfully. Without explanations, contestability may become superficial, as doctors would not have the information needed to identify errors in the system's logic【77†source】.
3. **Data Quality**: Another requirement of equal importance is **data quality**. If the data used to train AI systems is poor, even the best contestability mechanisms will not prevent erroneous recommendations. Ensuring high-quality, representative training data is a critical foundation for the reliability of AI systems.

In conclusion, while contestability is essential, **explainability** and **data quality** are equally important prerequisites for creating trustworthy AI systems that can be contested effectively. All three requirements must be prioritized in a clinical setting to ensure patient safety and high-quality care.