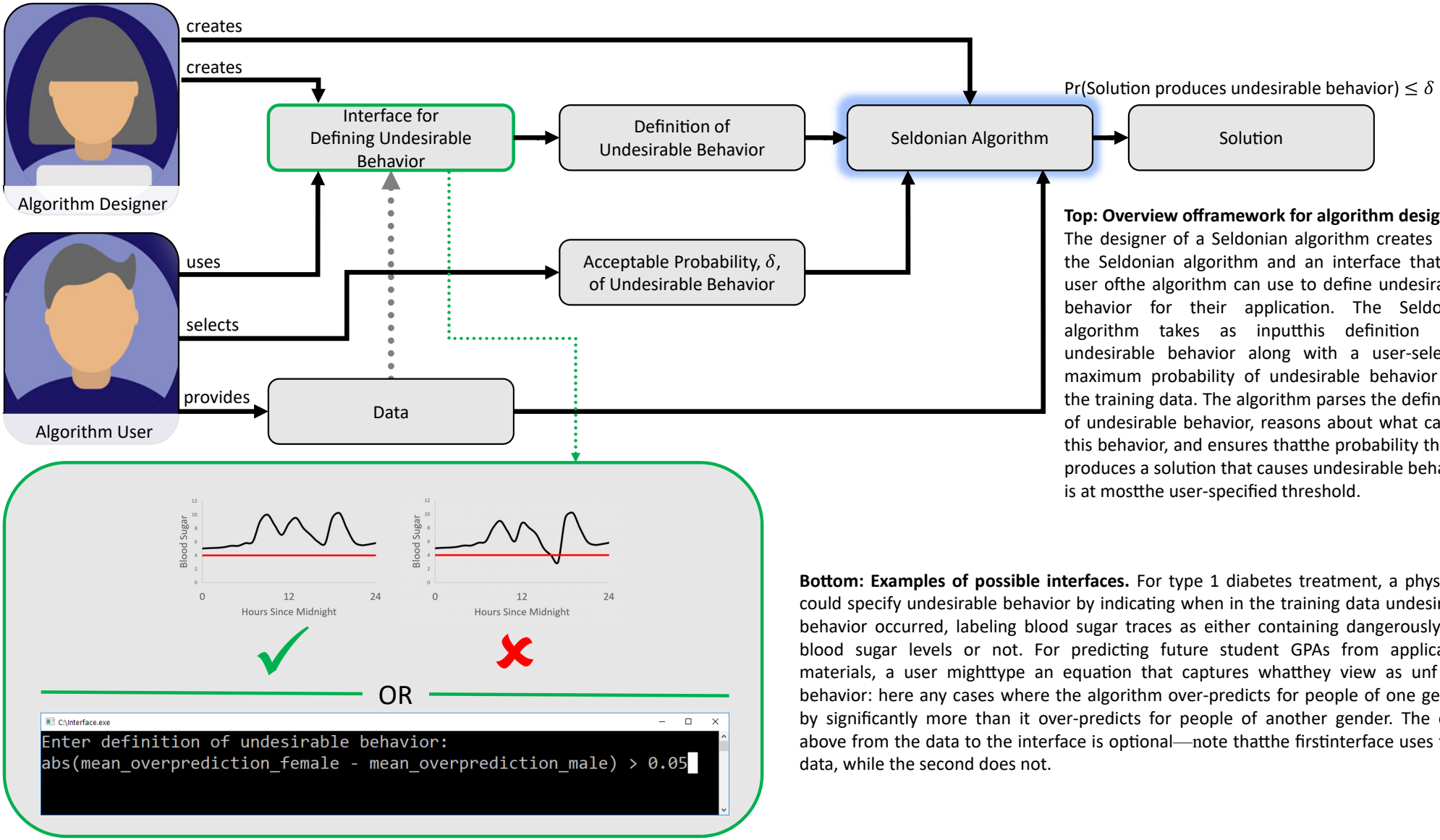

Top: Overview offramework for algorithm design. The designer of a Seldonian algorithm creates both the Seldonian algorithm and an interface thatthe user ofthe algorithm can use to define undesirable behavior for their application. The Seldonian algorithm takes as inputthis definition of undesirable behavior along with a user-selected maximum probability of undesirable behavior and the training data. The algorithm parses the definition of undesirable behavior, reasons about what causes this behavior, and ensures thatthe probability thatit produces a solution that causes undesirable behavior is at mostthe user-specified threshold.

Bottom: Examples of possible interfaces. For type 1 diabetes treatment, a physician could specify undesirable behavior by indicating when in the training data undesirable behavior occurred, labeling blood sugar traces as either containing dangerously low blood sugar levels or not. For predicting future student GPAs from application materials, a user mighttype an equation that captures whatthey view as unf air behavior: here any cases where the algorithm over-predicts for people of one gender by significantly more than it over-predicts for people of another gender. The edge above from the data to the interface is optional—note thatthe firstinterface uses the data, while the second does not.