Understanding Deep Learning

Chapter 21: Deep learning and ethics

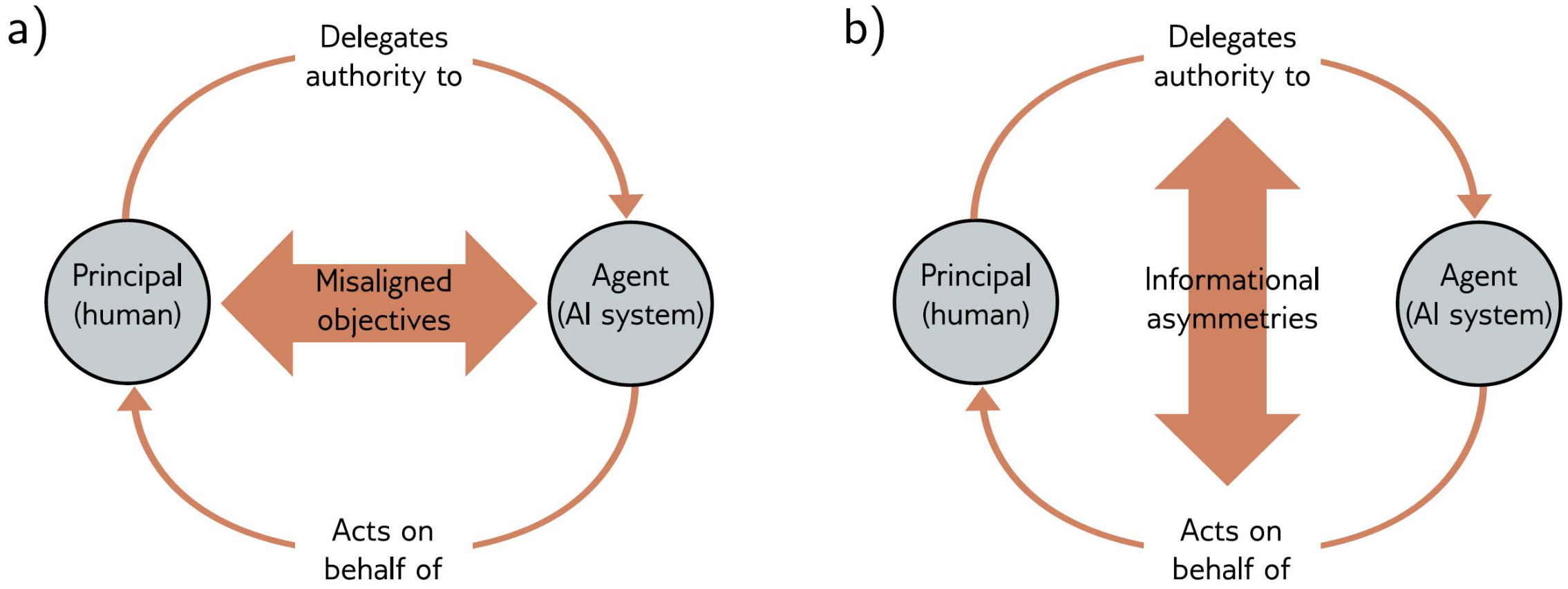

Figure 21.1 Structural description of the value alignment problem. a) Problems arise from a) misaligned objectives (e.g., bias) or b) informational asymmetries between a (human) principal and an (artificial) agent (e.g., lack of explainability). Adapted from LaCroix (2023).

Data collection	Pre-processing	Training	Post-processing
• Identify lack of examples or variates and collect	• Modify labels • Modify input data • Modify input/ output pairs	• Adversarial training • Regularize for fairness • Constrain to be fair	• Change thresholds • Trade-off accuracy for fairness

Figure 21.2 Bias mitigation. Methods have been proposed to compensate for bias at all stages of the training pipeline, from data collection to post-processing of already trained models. See Barocas et al. (2023) and Mehrabi et al. (2022).

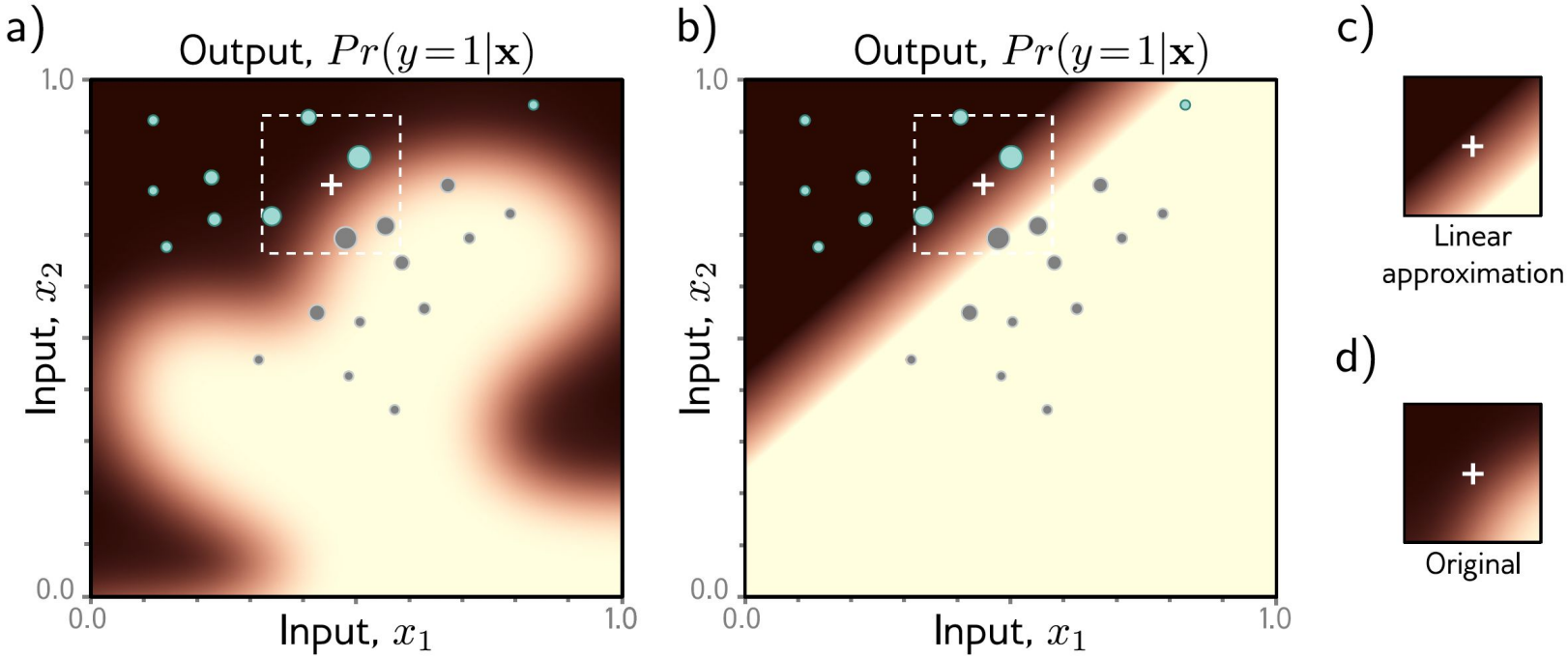

Figure 21.3 LIME. Output functions of deep networks are complex; in high dimensions, it's hard to know why a decision was made or how to modify the inputs to change it without access to the model. a) Consider trying to understand why $Pr(y = 1|\mathbf{x})$ is low at the white cross. LIME probes the network at nearby points to see if it identifies these as $Pr(y = 1|\mathbf{x}) < 0.5$ (cyan points) or $Pr(y = 1|\mathbf{x}) \geq 0.5$ (gray points). It weights these points by proximity to the point of interest (weight indicated by circle size). b) The weighted points are used to train a simpler model (here, logistic regression — a linear function passed through a sigmoid). c) Near the white cross, this approximation is close to d) the original function. Even though we did not have access to the original model, we can deduce from the parameters of this approximate model, that if we increase x_1 or decrease x_2, $Pr(y = 1|\mathbf{x})$ will increase, and the output class will change. Adapted from Prince (2022).