Abstract: Modern machine learning approaches to classification, including AdaBoost, support vector machines, and deep neural networks, utilize surrogate loss techniques to circumvent the computational complexity of minimizing empirical classification risk. These techniques are also useful for causal policy learning problems, since estimation of individualized treatment rules can be cast as a weighted (cost-sensitive) classification problem. Consistency of the surrogate loss approaches studied in Zhang (2004) and Bartlett et al. (2006) crucially relies on the assumption of correct specification, meaning that the specified set of classifiers is rich enough to contain a first-best classifier. This assumption is, however, less credible when the set of classifiers is constrained by interpretability or fairness, leaving the applicability of surrogate loss based algorithms unknown in such second-best scenarios. This paper studies consistency of surrogate loss procedures under a constrained set of classifiers without assuming correct specification. We show that in the setting where the constraint restricts the classifier's prediction set only, hinge losses (i.e., ℓ1-support vector machines) are the only surrogate losses that preserve consistency in second-best scenarios. If the constraint additionally restricts the functional form of the classifier, consistency of a surrogate loss approach is not guaranteed even with hinge loss. We therefore characterize conditions for the constrained set of classifiers that can guarantee consistency of hinge risk minimizing classifiers. Exploiting our theoretical results, we develop robust and computationally attractive hinge loss based procedures for a monotone classification problem.