Mitigating Moral Hazard in Cyber Insurance Using Risk Preference Design

Shutian Liu and Quanyan Zhu

Abstract—Cyber insurance is a risk-sharing mechanism that can improve cyber-physical systems (CPS) security and resilience. The risk preference of the insured plays an important role in cyber insurance markets. With the advances in information technologies, it can be reshaped through nudging, marketing, or other types of information campaigns. In this paper, we propose a framework of risk preference design for a class of principal-agent cyber insurance problems. It creates an additional dimension of freedom for the insurer for designing incentive-compatible and welfare-maximizing cyber insurance contracts. Furthermore, this approach enables a quantitative approach to reduce the moral hazard that arises from information asymmetry between the insured and the insurer. We characterize the conditions under which the optimal contract is monotone in the outcome. This justifies the feasibility of linear contracts in practice. This work establishes a metric to quantify the intensity of moral hazard and create a theoretical underpinning for controlling moral hazard through risk preference design. We use a linear contract case study to show numerical results and demonstrate its role in strengthening CPS security.

I. INTRODUCTION

Recent years have witnessed significant advances in the development and social impacts of Cyber-Physical System (CPS) technologies, including smart grids, autonomous vehicle systems, and implantable medical devices [1]. CPS are complex systems integrating various components and technologies (e.g., Information Technologies and Operational Technologies) to enable mission-critical functions. As a result, CPS inherits vulnerabilities from many sources. These vulnerabilities can be exploited by attackers and inflict CPS users significant losses, and they are hard to eliminate. There is a crucial need to reduce the risk of CPS users and increase the resiliency in face of attacks.

Cyber insurance is a financial product that transfers cyber risks between the insurer and the insured. Complementary to cyber protection/defense mechanisms, it aims to improve the cyber resiliency of a system by mitigating the financial losses as a result of successful attacks. Cyber insurance involves a contractual agreement that determines the premium and the coverage of the insurance contract. The premium is an amount of fee paid by the insured to the insurer to participate in the insurance program. The coverage specifies the amount of loss the insurer will cover. With a properly designed contract, the security of CPS users can benefit from cyber insurance.

One design framework for cyber insurance relies on a class of principal-agent models [2], which capture the information asymmetry between the insurer and the insured. The asymmetry leads to the moral hazard phenomenon [3] in which the insured users tend to behave more recklessly when interacting with the system and result in severe consequences. Risk aversion plays an important role in moral hazard. It is known that risk-averse behaviors of the user can increase the gap between the principal’s profits obtained from the contract with the full observability of the user’s action and the contract when the user’s action is hidden. This phenomenon is absent when the insurer meets a risk-neutral user [4]. Hence, measuring the level of risk aversion is essential in investigating the effect of moral hazard on cyber insurance.

Risk measure [5] provides a convenient way to quantify the risks and model agents’ risk perception or preferences. Different from the risk described under the expected utility theory, risk measure is richer in capturing individual perceptions of risks induced by the probabilistic nature of random outcomes. The design of risk measure is a way to control the risk preferences of the agents using exogenous influences such as nudging, marketing, and propaganda [6], [7].

This work aims to propose, as shown in Fig. 1, a risk preference design framework to shape the users’ risk preferences for high-performance cyber insurance. We consider a class of principle-agent problems in which one principal meets an idiosyncratic individual from a population of agents. Each agent has a risk preference type associated with an idiosyncratic risk measure. The goal of the design aims to optimize the profit of the insurer. We formulate the problem as a bilevel optimization problem. In the upper-level insurer’s problem, the insurer optimizes her profit by offering a contract and performing risk preference design subject to participation constraints. In the lower-level user’s problem, the user follows the claimed contract under the chosen risk preference.

Using the risk preference design framework, we characterize the solution to the insurer’s problem by imposing the conditions under which the optimal contract exhibits monotonicity. This paves the way for considering linear...
contracts in practice. We quantify the intensity of Moral Hazard (IMH) by benchmarking the action of the user from solving his problem with the action obtained from the full-information counterpart of the insurer’s problem. Leveraging the optimality conditions of the contract problems, we quantify how IMH varies according to the change in the risk preference type distributions. This analysis enables the risk preference design for mitigating moral hazard and creates a revenue-compatible design for the principal. In the case study, we use linear contracts as an example to show that risk preference design reduces moral hazard and improves security.

Section II provides a brief literature review. Section III introduces risk preference types and formulates the design problem. Section IV analyzes the optimal contract and the proposed approach to quantify moral hazard. Section V contains a case study. Section VI concludes the paper.

II. RELATED WORKS

One of the classic approaches to model risk perception relies on the expected utility theory of [8], where decisions of the players or the agents are assumed to maximize their utility. The concept is extended in [9] and [10] to the well-known Arrow-Pratt measure of absolute risk aversion. Another related approach is the cumulative prospect theory of [11], where the probability of the random outcomes is weighted. Motivated by the cumulative prospect theory, we use risk measures to capture individual risk preferences. The seminal work [5] has proposed to adopt coherent risk measures (CRMs) for risk quantification. The axioms that a CRM satisfies not only make it meaningful in practice but also enables the dual representation [5], [12]. This representation shows the robustness of a CRM and connects it with the popular distributionally robust optimization [13].

Risk design is closely related to system robustness [14]. Risk design in cyber insurance elicits human behaviors that have an impact on cyber security and resilience. In [15], the authors have proposed a pre-screening method to increase network security by designing incentive contract which increases users’ effort levels. In [16], the authors have have shown that an increased level of risk aversion increases the social welfare gain from using cyber insurance for protection. This work aims to consolidate the design of risk preferences of users into cyber insurance to further mitigate cyber risks.

III. PROBLEM FORMULATION

In this section, we introduce the framework of cyber insurance with risk preference design. We present the concept of risk preference types to model different perceptions of risks in connection with a class of principal-agent problems.

A. Risk preference types

Consider a mass 1 population of users of networked systems. This population faces uncertainties that arise from cyber risks captured by the probability space $(\Xi,\mathcal{F})$ with the reference probability measure $P$. We use $\xi \in \Xi \subset \mathbb{R}$ to denote a sampled outcome. We denote by $Z : \Xi \to \mathbb{R}$ a random loss, where $Z$ is a measurable function with finite $p$-th order moment from the space $\mathcal{L}^p := \mathcal{L}_p(\Xi,\mathcal{F},P)$.

$$\mathcal{L}_p := \mathcal{L}_p(\Xi,\mathcal{F},P)$$

The parameter $p$ lives in $[1, +\infty)$. Individuals in this population possess different risk preferences toward uncertainty. In particular, each individual is identified with a risk preference type $\theta \in \Theta \subset \mathbb{R}$. A risk preference of type $\theta$ determines the way user’s perception of risks, and it is captured by a risk measure $\rho_\theta : \mathcal{L}_p \to \mathbb{R}$.

Let $(\Theta,\mathcal{F})$ denote the underlying probability space of risk preference types. The risk preference types in the population are distributed according to a probability measure $\mu^0$ from the set of probability measures $\mathcal{D}$ defined on $(\Theta,\mathcal{F})$. Note that, in practice, $\mu^0$ can be known by collecting information through questionnaires or analyzing historical customer behaviors.

The goal of risk preference design is to find a distribution $\mu \in \mathcal{D}$ so that incentive contracts can be created between the users and the insurer to cover potential losses due to users’ cyber risks. We consider designing the distribution of risk preferences owing to the following three reasons. First, the utility of the insurer is measured by the outcome of a population of users participating in the contract. Hence the expected utility captures the incentives of the insurer. Second, it is difficult to design customized contracts as precise individual user’s risk perception is hard to measure. The collective pattern of the users of the same type can be statistically measured. Third, this distributional perspective can be endowed with an interpretation that the insurer and the same user interact repeatedly for a sufficiently long period of time. We will elaborate this point further in Section III-B.

We assume that the risk measures $\rho_\theta$ for all $\theta \in \Theta$ are CRMs [5]. Below is the definition of coherent risk measures.

**Definition 1 (Coherent Risk Measures [5]):** A function $\rho : \mathcal{L}_p \to \mathbb{R}$ is called a Coherent Risk Measure if it satisfies

(A1) Monotonicity: $\rho(Z) \geq \rho(Z')$ if $Z, Z' \in \mathcal{L}_p$ and $Z \geq Z'$.

(A2) Convexity: $\rho(tZ + (1-t)Z') \leq t\rho(Z) + (1-t)\rho(Z')$ for all $Z, Z' \in \mathcal{L}_p$ and $t \in [0,1]$.

(A3) Translation equivariance: If $Z \in \mathcal{L}_p$ and $a \in \mathbb{R}$, then $\rho(Z + a) = \rho(Z) + a$.

(A4) Positive homogeneity: $\rho(tZ) = t\rho(Z)$ if $Z \in \mathcal{L}_p$ and $t \in \mathbb{R}_+$.

One of the most significant consequences of CRMs is their dual representation [5], [12]. Let $\mathcal{D}^* := \mathcal{D}^*(\Theta,\mathcal{F},P)$ for $q \in (1,\infty)$ denote the dual space of $\mathcal{D}$, i.e., $1 \over p + 1 \over q = 1$. Following [5], [12], the dual representation of $\rho_\theta$ is

$$\rho_\theta[Z(\xi)] := \sup_{\xi \in \Xi} \int Z(\xi)dP(\xi, x),$$

where $\mathcal{A}_\theta \subset \mathcal{D}^*$ denotes the dual set associated with the risk measure $\rho_\theta$. The set $\mathcal{A}_\theta$ is convex and compact when $p \in [1, +\infty)$ [17]. Hence, the maximum of (1) is attained.

B. Cyber Insurance with Risk Preference Design

We incorporate the risk preference types discussed in Section III-A into a principal-agent problem by specifying the following action-outcome relation. Consider a principal-agent problem with one principal and one agent. Let $x \in X$ denote the agent’s action, which represents the amount of security...
investment that he spends on his device. We set \( \Xi := [\xi_1, \xi_2] \) and identify it as the set of random cyber risks which affect the costs of both the principal and the agent. The parameterized probability measure \( P(\xi, x) \) describes the stochastic relation between the cyber risk and the security investment. We assume that for all \( x \in X, P(\xi, x) \) is absolutely continuous with the reference probability measure \( P(\xi), \) i.e., there is a density function \( p(\xi, x) \) such that \( dP(\xi, x) = p(\xi, x) dP(\xi). \) And we assume that \( p(\cdot, x) \) is differentiable for all \( x \in X \) and \( p(\cdot, \cdot) \) is continuously differentiable for all \( \xi \in \Xi. \)

The insurer acts as the principal in the problem. She has the power of claiming a coverage plan of the cyber losses described by the function \( w(\cdot): \Xi \to \mathbb{R} \) and choosing a distribution of risk preference types \( \mu \in \mathcal{D} \) for the users. For a given cyber risk \( \xi, \) she observes cost (or negative profit) \( w(\xi) + \xi. \) We assume that the insurer is risk-neutral. Hence, her expected cost from claiming a plan \( w(\cdot) \) is \( \int_{\Xi} w(\xi) + \xi dP(\xi, x). \) The manipulation of risk preferences can come with a cost. We introduce the order-1 Wasserstein distance \([18] W_1(\mu, \mu^0)\) to represent the cost associated with the effort to change the risk preference distribution from \( \mu^0 \) to \( \mu. \) The cost of the insurer in designing \( w(\cdot) \) and \( \mu \) is

\[
\int_{\Xi} w(\xi) + \xi dP(\xi, x) + \gamma W_1(\mu, \mu^0),
\]

where \( \gamma > 0 \) represents the monetary cost of risk manipulation. Consider a user with type \( \theta. \) If he takes action \( x_i \), under plan \( w(\cdot), \) the cyber risk that he perceives is \( \rho_\theta[U(w(\xi), x_i)] \) with the random cost \( U \in \mathcal{D}. \) We assume that \( U \) is increasing and convex in \( w, \) which indicates that the users exhibit loss aversion. Note that since \( U \) can be understood as a disutility function, our assumptions on \( U \) are consistent with the assumptions that a utility function is non-decreasing and concave. In the literature, these assumptions are often referred to as ‘risk aversion’ [2]. We use the term loss aversion to distinguish it from risk preferences represented by the risk measures \( \rho_\theta. \) Our treatment of risk is motivated by [11], where the authors combine the loss aversion and the probability weighting to characterize decisions under risk.

The agent is assumed to be an idiosyncratic individual from the population of users who bears the averaged action \( x \). When the type distribution is \( \mu, \) the cost of the agent is

\[
\int_{\Xi} \rho_\theta[U(w(\xi), x)] d\mu(\theta).
\]

Note that in (3), the randomness is addressed by the risk measures \( \rho_\theta; \) the integral over \( \Theta \) represents an averaging over all the risk preference which generates the averaged action \( x. \)

Let \( U > 0 \) denote the deterministic cyber risk threshold that the agent can bear. The hidden action cyber insurance problem with risk-preference design is formulated as follows:

\[
\begin{align*}
\min_{w(\cdot), \mu(\cdot), x, \xi} & \int_{\Xi} w(\xi) + \xi dP(\xi, x) + \gamma W_1(\mu, \mu^0) \\
\text{s.t.} & \int_{\Xi} \rho_\theta[U(w(\xi), x)] d\mu(\theta) \leq U, \quad (\text{IR}), \\
& x \in \arg\min_{x' \in X} \int_{\Xi} \rho_\theta[U(w(\xi), x')] d\mu(\theta), \quad (\text{IC}).
\end{align*}
\]

In (4), (IR) is the individual rationality constraint guaranteeing that participation benefits the agent; (IC) is the incentive compatibility constraint stating that the agent should minimize his own cost when his action is hidden.

The main difference between (I) and the classical principal-agent problems is that the distribution \( \mu \) characterizing the distribution of risk preferences of the users becomes part of the design, offering a new degree of freedom to model the informational influences on the users and create additional control to elicit proper cyber hygiene of the users and, at the meantime, improve the social welfare and the profit. Furthermore, as we will show in Section IV, the risk preference design mitigates the moral hazard issue.

IV. ANALYSIS OF THE DESIGN FRAMEWORK

A. Characterization of Optimal Contracts

In this section, we use the first-order approach to characterize several properties of the optimal coverage plan for the contract problems defined in Section III. We obtain conditions for the monotonicity of \( w(\cdot). \)

Consider the first-order optimality condition of (IC):

\[
\frac{\partial}{\partial \theta} \int_{\Xi} \rho_\theta[U(w(\xi), x)] d\mu(\theta) = 0.
\]

We assume that the set of density functions which attains the maximum of (I) is a singleton for all \( \theta \in \Theta, \) i.e.,

\[
\partial \rho_\theta[U(w(\xi), x)] = \{\xi_0(\xi, x)\}.
\]

Then, (5) leads to

\[
0 = \frac{\partial}{\partial \xi} \int_{\Xi} U(w(\xi), x) \xi_0(\xi, x) p(\xi, x) dP(\xi) d\mu(\theta)
\]

\[
= \frac{\partial}{\partial \xi} \int_{\Xi} \int_{\Xi} \rho_\theta[U(w(\xi), x)] d\mu(\theta) U(w(\xi), x) p(\xi, x) dP(\xi).
\]

Ignoring the second-order optimality condition of (IC), we find (6) as an alternative for the (IC) in (4).

The Lagrangian of problem (4) when (IC) is substituted by the first-order condition (6) can be expressed as the following:

\[
\mathcal{L} = \int_{\Xi} w(\xi) + \xi dP(\xi, x) + \gamma W_1(\mu, \mu^0) + \alpha \int_{\Xi} \rho_\theta[U(w(\xi), x)] d\mu(\theta) - U + \beta \frac{\partial}{\partial x} \mathbb{E}_\mu[\xi_0(\xi, x) U(w(\xi), x) p(\xi, x) dP(\xi)] d\mu(\theta).
\]

Define \( \pi(\xi, \theta, x) = \xi_0(\xi, x) p(\xi, x). \) Minimizing \( w(\xi) \) pointwise in \( \xi, \) we obtain the first-order condition of (7) for \( \xi \in \Xi: \)

\[
- p(\xi, x) = \mathbb{E}_\mu \left\{ \alpha \pi(\xi, \theta, x) \frac{\partial}{\partial w} U(w(\xi), x) + \beta \frac{\partial}{\partial x} \mathbb{E}_\mu[\xi_0(\xi, x) U(w(\xi), x) p(\xi, x) dP(\xi)] d(\pi(\xi, \theta, x)) \right\}.
\]

Before we proceed, we make the assumption that

\[
\frac{\partial^2}{\partial w^2} U(w(\xi), x) = 0.
\]

Note that this assumption is weaker than the one that the agent’s cost is separable in the security investment and the gain from the coverage plan in the cyber insurance contract. After several steps of algebraic manipulations, (8) leads to

\[
- p(\xi, x) = \mathbb{E}_\mu \left\{ \alpha \pi(\xi, \theta, x) \frac{\partial}{\partial w} U(w(\xi), x) + \beta \frac{\partial}{\partial x} \mathbb{E}_\mu[\xi_0(\xi, x) U(w(\xi), x) p(\xi, x) dP(\xi)] d(\pi(\xi, \theta, x)) \right\}.
\]
Dividing both sides of (9) by $p(\xi, x)$, we obtain

$$-\frac{1}{\sigma} \nabla_x U(w(\xi), x)$$

$$= \mathbb{E}_\mu \left\{ \alpha \xi_0(\xi, x) + \beta \left( \frac{\partial}{\partial x} \xi_0(\xi, x) + \xi_0(\xi, x) \frac{\partial^2 p(\xi, x)}{\partial \xi^2} \right) \right\}$$

$$= \mathbb{E}_\mu \left\{ \xi_0(\xi, x) \right\} \cdot \left( \alpha + \beta \frac{\partial^2 p(\xi, x)}{\partial \xi^2} \right) + \mathbb{E}_\mu \left\{ \frac{\partial}{\partial x} \xi_0(\xi, x) \right\}.$$

Differentiating with respect to $\xi$ over $\mathcal{Z}$ on both sides of (10), we obtain

$$w'(\xi) - \frac{\partial}{\partial \xi} U(w(\xi), x) \frac{\partial^2 U}{\partial \xi^2} = \beta \frac{\partial}{\partial \xi} \mathbb{E}_\mu \left\{ \frac{\partial}{\partial \xi} \xi_0(\xi, x) \right\}$$

$$+ \mathbb{E}_\mu \left\{ \xi_0(\xi, x) \right\} \cdot \left( \alpha + \beta \frac{\partial^2 p(\xi, x)}{\partial \xi^2} \right).$$

The above discussions lead to the following result, which characterizes the monotonicity of the optimal coverage plan.

**Proposition 1:** Suppose that the density function $\xi_0(\cdot, x)$ is differentiable on a subset $\mathcal{Z} \subset \mathcal{X}$. Then, the optimal coverage plan $w$ of (4) is increasing in the cyber risk on $\mathcal{Z}$, i.e., $w'(\xi) > 0$, if the following conditions hold:

1. (C1) The functions $p(\xi, x)$ and $\xi_0(\xi, x)$ satisfy $\frac{\partial^2}{\partial \xi^2} p(\xi, x) > 0$ and $\frac{\partial^2}{\partial \xi^2} \xi_0(\xi, x) > 0$, respectively;
2. (C2) Severe loss avoidance:

$$\frac{\partial^2}{\partial \xi^2} \mathbb{E}_\mu \left\{ \frac{\partial}{\partial \xi} \xi_0(\xi, x) \right\} \geq 0;$$

3. (C3) Risk-sensitive monotone likelihood ratio property:

$$\frac{\ partial}{\partial \xi} \mathbb{E}_\mu \left\{ \alpha + \beta \frac{\partial^2 p(\xi, x)}{\partial \xi^2} \right\} \geq 0.$$

**Proof:** By combining (C1) with the convexity of $U$ with respect to $w$, we can show $\beta > 0$ using the similar arguments as in [3]. When (C2) and (C3) hold, the optimality condition (11) leads to the monotonicity of $w$.

The conditions in Proposition 1 have the following interpretations. Condition (C1) implies two facts. First, without taking the user’s risk preference into account, the true likelihood for higher cyber risks to occur decreases when the security investment increases. Second, the perceived likelihood for higher cyber risks to occur of a user with risk preference type $\theta$ is decreased when the security investment increases. These two facts are consistent with the cyber insurance context. Before we elaborate on condition (C2), we first observe the following fact. Since the coverage plan $w$ is increasing in the cyber risk $\xi$ and the cost $U$ is increasing in $w$, we conclude that $U$ is also increasing in $\xi$. In addition, from (11), we know that for all $\theta$ the optimal density function $\xi_0(\xi, x)$ is increasing in $\xi$, i.e., $\frac{\partial^2}{\partial \xi^2} \xi_0(\xi, x) > 0$. Then, condition (C2) means that an increase in the user’s security investment shows his panic towards the potential occurrences of severer cyber risks. In other words, the user tries to avoid severe losses. Condition (C3) is the risk-sensitive counterpart of the monotone likelihood ratio property:

$$\frac{\partial}{\partial \xi} \mathbb{E}_\mu \left\{ \alpha + \beta \frac{\partial^2 p(\xi, x)}{\partial \xi^2} \right\} \geq 0.$$

Note that the conditions for proving monotonicity on $\mathcal{Z}$ can reduce to simply (C1) and the standard monotone likelihood ratio property for some commonly used CRMs, such as the mean semideviation measure and the average value-at-risk measure [12]. The reason lies in that the density function $\xi_0(\xi, x)$ of these two measures is piecewise constant. It means that the right-hand side of (11) becomes

$$\mathbb{E}_\mu \left\{ \xi_0(\xi, x) \right\} \cdot \left( \alpha + \beta \frac{\partial^2 p(\xi, x)}{\partial \xi^2} \right).$$

Then, we can recover [12].

**B. Mitigation of moral hazard**

It is well-known from the literature that one of the causes of moral hazard is the issue of hidden action; i.e., the insurer cannot observe the action taken by the insured. In this section, we propose to measure the IMH and provide a method to mitigate moral hazard. For notational simplicity, we consider an $n$-dimensional finite risk preference type space $\Theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$ with probability measure $\mu = (\mu_1, \mu_2, \ldots, \mu_n)^T$.

**a) Measuring the IMH:** We first consider the following full-information benchmark associated with (4) assuming that the risk preference distribution $\mu$ is given:

$$\min_{w(\xi), x} \int \xi dP(\xi, x) + \gamma W_1(\mu, \mu^0)$$

s.t. $\sum_{i=1}^n \rho_0 \left[ U(w(\xi), x) \right] \cdot \mu_i \leq \mathcal{U}$. (IR)

Let $w^*$ and $x^*$ denote the optimal solution of (13) given $\mu$. The Lagrangian of (13) is

$$L_\theta = \int \left( \xi dP(\xi, x) + \gamma W_1(\mu, \mu^0) \right)$$

$$+ \alpha \sum_{i=1}^n \rho_0 \left[ U(w(\xi), x) \right] \cdot \mu_i.$$

Then, the action $x^*$ solves

$$0 = \frac{\partial}{\partial x} L_\theta = \frac{\partial}{\partial x} \int \left( w^*(\xi) + \xi dP(\xi) \right) \cdot \mu.$$

Hidden actions refer to the fact that $x^*$ will not be the real action taken by the agent after receiving $w^*$ announced by the principal. From the agent’s perspective, an action $x^*$ optimizing his own objective is superior, i.e., $x^*$ such that

$$x^* \in \arg \min_{x \in \mathcal{X}} \sum_{i=1}^n \rho_0 \left[ U(w^*(\xi), x) \right] \cdot \mu_i.$$

Therefore, the deviation from $x^*$ to $x^*$ is the result of the moral hazard issue and $x^* - x^*$ is a meaningful metric for measuring the IMH. The value of $x^* - x^*$ is nonnegative because of the assumptions made in Section III.

**Definition 2:** We refer to the quantity $x^* - x^*$ as the IMH. A high value of $x^* - x^*$ indicates an intense moral hazard.

Note that (16) is part of the IC constraint in (4). However, the actions solving (4) and (16) are different in general.
b) Mitigation of moral hazard through risk preference design: We study how IMH varies according to the changes in \( \mu \). From (16), \( x' \) satisfies the following first-order condition:

\[
\frac{\partial}{\partial x} \sum_{i=1}^{n} \rho_i [U(w(\xi_{i}, x))] \cdot \mu_i = 0.
\] (17)

Define \( H_1 : X \times Q \to \mathbb{R} \) and \( H_2 : X \times Q \to \mathbb{R} \) according to

\[
H_1(x, \mu) = \frac{\partial}{\partial x} \sum_{i=1}^{n} \rho_i [U(w(\xi_{i}, x))] \cdot \mu_i.
\]

as in (15), and

\[
H_2(x, \mu) = \frac{\partial}{\partial x} \sum_{i=1}^{n} \rho_i [U(w(\xi_{i}, x))] \cdot \mu_i.
\] (19)
as in (17), respectively. Then, we obtain from (15) and (17) that, for a given \( \mu \), \( H_1(x', \mu) = 0 \) and \( H_2(x', \mu) = 0 \).

The next result states how to reduce IMH by designing the distribution of risk preference types.

Proposition 2: Let \( x' \) and \( x'' \) denote the isolated local minima of (15) and (16) given \( \mu \), respectively. Let \( J_{H_1,x} \) denote the Jacobian matrix of \( H_1 \) with respect to the first argument; let \( J_{H_2,x} \) denote the Jacobian matrix of \( H_2 \) with respect to the first argument. Then, in a neighborhood \( \mathcal{N}_{\mu} \subset \mathbb{R} \) of \( \mu \), the IMH can be reduced by changing \( \mu \) to \( \mu' = \mu + c\Delta \mu \) with \( c > 0 \) sufficiently small, if \( \Delta \mu \) satisfies

\[
(\Delta \mu)^T \cdot \nabla \mu \leq 0,
\]

where \( \nabla \mu = (\frac{\partial}{\partial \mu_i} \mu_1, \frac{\partial}{\partial \mu_i} \mu_2, \ldots, \frac{\partial}{\partial \mu_i} \mu_n)^T \) with

\[
\frac{\partial}{\partial \mu_i} \mu_j = (J_{H_1,x}(x', \mu))^{-1} \frac{\partial H_1}{\partial \mu_i}(x', \mu) - (J_{H_2,x}(x', \mu))^{-1} \frac{\partial H_2}{\partial \mu_i}(x', \mu),
\]

for all \( i = 1, 2, \ldots, n \).

Proof: Consider the optimization problems (13) and (16).

If \( H_1(x', \mu) = 0 \) and \( H_2(x', \mu) = 0 \). Since the local minima \( x' \) and \( x'' \) are isolated, the second-order conditions require that \( \frac{\partial^2}{\partial x^2} H_1 > 0 \) and \( \frac{\partial^2}{\partial x^2} H_2 > 0 \). Equivalently, we obtain \( J_{H_1,x}(x', \mu) > 0 \) and \( J_{H_2,x}(x', \mu) > 0 \). Invoking the implicit function theorem, we know that there exist differentiable functions \( h_1(\cdot) \) and \( h_2(\cdot) \) defined on neighborhoods \( \gamma_{\mu} \times \gamma_{\mu} \) of \( (x', \mu) \) and \( \gamma_{\mu} \times \gamma_{\mu} \) of \( (x', \mu) \), such that \( h_1(\mu) = x' \) and \( h_2(\mu) = x'' \). Furthermore, the derivatives of the functions \( h_1(\cdot) \) and \( h_2(\cdot) \) with respect to \( \mu \) for all \( i = 1, 2, \ldots, n \) are

\[
\frac{\partial h_1}{\partial \mu_i}(\mu) = - (J_{H_1,x}(h_1(\mu), \mu))^{-1} \frac{\partial h_1}{\partial h_1}(h_1(\mu), \mu),
\]

and

\[
\frac{\partial h_2}{\partial \mu_i}(\mu) = - (J_{H_2,x}(h_2(\mu), \mu))^{-1} \frac{\partial h_2}{\partial h_2}(h_2(\mu), \mu),
\]

respectively. Let \( \mathcal{N}_{\mu} = \gamma_{\mu} \cap \gamma_{\mu} \). Define \( T(\mu) = h_1(\mu) - h_2(\mu) \) with support \( \mathcal{N}_{\mu} \). Then, the quantity \( x' - x'' \) can be represented using \( T(\mu) \) for \( \mu \in \mathcal{N}_{\mu} \). The gradient \( \nabla T(\mu) = (\frac{\partial T}{\partial \mu_1}(\mu), \frac{\partial T}{\partial \mu_2}(\mu), \ldots, \frac{\partial T}{\partial \mu_n}(\mu))^T \) follows from (22) and (23) for \( \mu \in \mathcal{N}_{\mu} \). Next, consider \( T(\mu') \), which represents the IMH when the distribution of risk preference is \( \mu' = \mu + c\Delta \mu \). Taylor’s theorem indicates that:

\[
T(\mu') = T(\mu) + c(\Delta \mu)^T \cdot \nabla T(\mu) + o(c).
\]

For sufficiently small \( c > 0 \), condition (20) leads to \( T(\mu') \leq T(\mu) \). This completes the proof.

Obtaining the value of IMH relies on solving problems (16) and (13). However, Proposition 2 offers a method to check whether a target distribution of risk preference types mitigates the effect of the moral hazard.

c) Design methods: We have shown that the IMH can be mitigated by risk preference design. Next, we show that the direction \( \Delta \mu \) in (20) can decrease the objective value of (13), making \( \mu' = \mu + c\Delta \mu \) a favorable design.

Before stating the results, we first present an equivalent representation of \( W_1(\mu, \mu') \). Since we consider finite probability measures, strong duality holds in the Kantorovich-Rubinstein dual problem (18) of \( W_1(\mu, \mu') \), i.e.,

\[
W_1(\mu, \mu') = \max_{b \in B} \sum_{i=1}^{n} b_i \mu_i - \sum_{i=1}^{n} \sum_{j=1}^{n} b_j \rho_j(\mu'_i - \mu_j),
\]

where \( B = \{ b \in \mathbb{R}^n : |b_i - b_j| \leq |i - j|, \forall i, j = 1, 2, \ldots, n \} \).

Proposition 3: For a given \( \mu \), let \( \nabla T(\mu) \) be defined as in Proposition 2. Let \( b^* \in \mathbb{R}^n \) denote the solution of (24). Then, there exists \( \Delta \mu \), such that by changing \( \mu \) to \( \mu' = \mu + c\Delta \mu \), the IMH and the objective cost of (13) decrease, if there does not exist \( a > 0 \), such that

\[
-ab^* = \nabla T(\mu).
\]

Proof: Since in the objective function of (13), the only term depending on \( \mu \) is \( W_1(\mu, \mu') \), we only need to show that \( W_1(\mu', \mu) \leq W_1(\mu, \mu') \) for \( \mu' = \mu + c\Delta \mu \). Equivalently, we show that \( \Delta \mu \) is a descent direction of \( W_1(\mu, \mu') \). To obtain the descent direction, we compute the gradient of \( W_1(\mu, \mu') \) using the representation (24). Since (24) is an optimization problem, the envelope theorem indicates that

\[
\nabla \mu W_1(\mu, \mu') = \nabla \mu \left( \sum_{i=1}^{n} b_i \mu_i - \sum_{i=1}^{n} \sum_{j=1}^{n} b_j \rho_j(\mu'_i - \mu_j) \right) = b^*,
\]

where \( b^* \) solves (24). Then, \( \Delta \mu \) descends \( W_1(\mu, \mu') \) if

\[
(\Delta \mu)^T \cdot b^* \leq 0.
\]

From Proposition 2 we know that changing \( \mu \) to \( \mu' \) mitigates the IMH if \( \Delta \mu \) satisfies condition (20). Hence (20) and (27) lead to the condition (25). This completes the proof.

The results in Proposition 2 and Proposition 3 lead to strengthened security for the following reasons. A lower value of the IMH indicates a smaller difference in the action ideal for the insurer and the action the user chooses in reality. It means that there is an increased likelihood to sign a contract where the agreed action generates a reasonable amount of cyber risk that the insurer can bear, and it is not overly challenging for the user to execute. Furthermore, to come up with such a contract, the insurer needs to prepare for the consumption required for shaping the users’ perceptions. Once the cost of risk design and the cost of coverage do not exceed the payment from the user, a contractual agreement which improves cyber security can be reached. We will elaborate on this point in Section VII.

V. CASE STUDY: LINEAR CONTRACTS

In this section, we use a special case study of linear contracts to show that risk preference design strengthens cyber security. We consider a scenario where an attacker

\[
T(\mu') = T(\mu) + c(\Delta \mu)^T \cdot \nabla T(\mu) + o(c).
\]
when the security investment increases, there is a lower
The cost for the user is \( \bar{C} \) when the system
distribution of the types a design parameter in the framework,
capture their perceptions of cyber losses. We have made the
contracts for cyber-physical systems security. We have defined
enables risk sharing between the user and the insurer but also
security effort from the user to protect the system. It not only
three possible losses due to the ransomware (\( \xi_1, \xi_2, \xi_3 \)) = 
(1,2,3). The probability of \( \xi(\xi_1, \xi_2, \xi_3) \) is (0.3,0.4,0.3) under \( x_L \) and (0.5,0.3,0.2) under \( x_H \). Note that this indicates that
when the security investment increases, there is a lower
likelihood for more severe losses to occur.

The insurer offers a linear contract containing a coverage
\( c \in (0,1) \) with a premium \( p > 0 \). Her expected cost from this
contract is \( E(\bar{C} - p) \). Since \( c \in (0,1) \), she prefers the high
investment \( x_H \) to the low investment \( x_L \) from the user. We
set the user’s loss aversion to be quadratic; i.e., he perceives
\( \xi^2 \) when the cyber loss is \( \xi \). The relative cost of investment
is \( m > 0 \). If the user does not buy insurance, he receives
the full cyber loss \( \xi \) without any coverage from the insurer.

The cost for the user is \( U(x) = \sum_{i=1}^{2} \mu_i \cdot \rho_0(q^2 + m) \). If he
purchases the insurance, the cost becomes \( \tilde{U}(x) = \sum_{i=1}^{2} \mu_i \cdot \rho_0([1 - c]q^2 + m + p) \).

Suppose \( m(x_H - x_L) > 1.1c^2 + 0.6\kappa\sigma^2 \xi^2 \). Then, \( \bar{U}(x_H) > \tilde{U}(x_L) \). From (IC) of \( H \), we know that the user will only agree
on \( x_L \) in a contract. The moral hazard issue occurs because
of the distinction between the insurer’s and user’s preferences
over the security investments. Note that participating in the
contract is profitable if \( p \leq (2c - c^2) \cdot (3.5 + 1.25\kappa\sigma^2) \).

With risk preference design, it is straightforward to observe that
as long as we increase the portion of type \( \theta_1 \) to \( \mu_1 \) such that
\( m(x_H - x_L) < 1.1c^2 + 0.6\kappa\sigma^2 \mu_1 \), the user prefers \( x_H \) to \( x_L \).

It means that by designing the distribution of risk preference
types so that more individuals evaluate risk according to the
absolute semideviation measure, the moral hazard issue can be
mitigated. One of the features of the absolute semideviation
measure is that it emphasizes the high impact outcomes in
spite of their low possibilities. Hence, a user adopting
this perception of risk shows aversion toward severe system
losses caused by ransomware. The consequence of this risk
preference design is that a reached contract elicits a high-level
security effort from the user to protect the system. It not only
enables risk sharing between the user and the insurer but also
enhances the overall defense against ransomware.

VI. CONCLUSION

In this paper, we have proposed cyber insurance with a
risk preference design framework to provide risk-sharing
contracts for cyber-physical systems security. We have defined
risk preference types for users and adopted risk measures to
capture their perceptions of cyber losses. We have made the
distribution of the types a design parameter in the framework,