Games **2013**, *4*, 508-531; doi:10.3390/g4030508

OPEN ACCESS

games

ISSN 2073-4336

www.mdpi.com/journal/games

Article

Solution Concepts of Principal-Agent Models with Unawareness of Actions

Ying-Ju Chen [1] and Xiaojian Zhao [2,]*

[1] University of California at Berkeley, 4121 Etcheverry Hall, Berkeley, CA 94720, USA;
E-Mail: chen@ieor.berkeley.edu

[2] Department of Economics, Hong Kong University of Science and Technology, Hong Kong;
E-Mail: xjzhao@ust.hk

* Author to whom correspondence should be addressed; Email: xjzhao@ust.hk, Tel.: +852-2358-7610.

Received: 14 February 2013; in revised form: 12 August 2013 / Accepted: 19 August 2013 / Published: 30 August 2013

Abstract: In numerous economic scenarios, contracting parties may not have a clear picture of all the relevant aspects. A contracting party may be unaware of what she and/or others are entitled to determine. Therefore, she may reject a contract that is too good to be true. Further, a contracting party may actively exert cognitive effort before signing a contract, so as to avoid being trapped into the contractual agreement *ex post*. In this paper, we propose a general framework to investigate these strategic interactions with unawareness, reasoning and cognition and intend to unify the solution concepts in the contracting context with unawareness. We build our conceptual framework upon the classical principal-agent relationship and compare the behaviors under various degrees of the unaware agent's sophistication.

Keywords: unawareness; cognition; incomplete contracts; principal-agent relationship

1. Introduction

In numerous economic scenarios, contracting parties may not be aware of all aspects of relevant actions. As a canonical example, consider the employee-employer contracting scenario. An employee may be unaware of the possibility of obtaining some training to improve her productivity and may not know *ex ante* that the employer could provide a poor retirement plan. Housing subsidies, childcare backup programs, additional research grant opportunities, dual-career consultation and expected administrative and committee work all may not be explicitly written in the contract. While confronted with these unawareness issues and the potential exploitation by others, the strategic decisions of the contracting parties critically depend on their *sophistication*. A naive contracting party may take the contract offer as given and passively expand her view of the world to include all the terms in the contract of which she was previously unaware. A more sophisticated contracting party may attempt to put herself in the other party's shoes to evaluate whether a proposed contract is a honest mutually beneficial deal, a sloppy contract offered by a error-prone party or a trap intentionally set up to take advantage of her. Further, a contracting party may actively exert cognitive effort before signing a contract in order to compensate/overcome the asymmetric awareness.

When a contractual relationship involves such unawareness, reasoning and cognitive effort, the optimal contract design (from the contract proposer's perspective) becomes subtle. On the one hand, since the contract follower (hereafter the *agent*) is not fully aware of all the aspects relevant to the contractual relationship, the contract proposer (hereafter the *principal*) may strategically disclose only a subset of relevant aspects in the contract at his own benefit. On the other hand, the intentionally concealed information may make a sophisticated agent suspect that something may go wrong and take some defensive counteraction, such as refusing the contract or actively gathering information. These inherent economic trade-offs give rise to a number of interesting issues. Given a contract offer, how does an unaware agent update her information? How does an agent rationalize the principal's contract offer? If a contract offer is not reasonable, how does the agent perceive and respond? How should the principal design the optimal contracts based on the agent's sophistication?

To address these issues, we construct a fairly general principal-agent model. As is standard in the principal-agent literature, we assume that the agent's actions are not *observable*, whereas all actions of the principal are *verifiable*. However, the agent may be *unaware* of all the relevant aspects of the actions. This unawareness is modeled by introducing the missing dimensions of the strategy set, which is akin to the unawareness models of missing dimensions of the full state space (see [1–3]). By way of contrast, the principal is fully aware of the entire strategy sets of both the principal and the agent and knows the agent's awareness. Since the agent may be unaware, the principal can determine whether to inform the agent via the contract offers. This contract offer may serve as an *eye-opener* that broadens the agent's vision and allows the agent to get a better understanding of the entire picture.

Based on the framework above, we propose a number of *solution concepts* that account for various degrees of the unaware agent's sophistication. As a direct extension of the classical subgame perfect Nash equilibrium to incorporate the agent's unawareness, we first introduce the *subgame-perfect solution* in which the agent updates her unawareness based on the principal's contract offer. The novel feature that arises from the agent's unawareness is that there is room for the principal to determine what to announce/include in the contract and which actions to implement in the aspects not specified in the contract. Since the principal and the agent perceive different games, the principal's contract offer may not be optimal from the agent's viewpoint. This is in strict contrast with the standard game theory that assumes common knowledge of the game. This discrepancy creates room for various choices of alternative solution concepts, as we elaborate below.

The second solution concept we introduce is the *justifiable solution*. Under this solution concept, if based on the agent's investigation, the principal should have offered an alternative contract, the agent suspects that something has gone wrong and, therefore, may reject the contract to avoid the potential exploitation. The agent's reasoning upon receiving a contract alters what the principal is able to offer, thereby giving rise to an additional "justifiability" constraint on the principal's side. The justifiable solution is intended to capture the idea that an unaware agent may still be able to evaluate whether the principal's contract offer is "reasonable" (see [4,5] and, more fundamentally, [6]). It is also similar to that of *forward induction* in game theory, as the subsequent player also reasons the former player's motivation upon observing the former player's actions ([7,8]).

In the third solution concept, we intend to capture the idea that while confronted with a non-justifiable contract, the agent believes that this contract may be unintended and simply result from *the principal's mistake*.[1] In such a scenario, we can conveniently assume that from the agent's perspective, a non-justifiable contract results from the principal's mistake with probability $1 - \rho$, and with probability ρ, this non-justifiable contract is a trap set up by the principal. With these probabilities, the agent then decides whether to accept the contract based on her expected utility, which leads to a *trap-filtered solution*. Note that when $\rho = 0$, the agent is extremely confident that any non-justifiable contract should be attributed to the principal's mistake, and the trap-filtered solution degenerates to a subgame-perfect solution. On the other hand, if $\rho = 1$, whenever she sees a non-justifiable contract, she perceives it as a trap, and the trap-filtered solution coincides with the justifiable solution. Thus, the trap-filtered solution can be regarded as a broader family of the solution concepts, and it nicely synthesizes all possible scenarios regarding how the agent perceives the principal's contract offer.

Finally, we investigate the scenario in which the agent is able to exert some effort to revise the belief about ρ upon receiving a non-justifiable contract. This *cognitive effort* allows the agent to pull back from being trapped into an intentional non-justifiable contract with the principal. As in [10], such cognitive effort is definitely helpful for the agent, but it comes at a cost. The higher cognitive effort the agent spends *ex ante*, the more likely she is able to identify a contractual trap given that there is indeed a trap. Thus, the principal must take into account the agent's cognitive effort and the possible consequences upon designing the contract.[2]

Our main contribution is to provide a general framework that synthesizes a number of seemingly unrelated solution concepts in [4,10–12], which allows us to investigate the interactions among unawareness, reasoning and cognitive effort in the optimal contract design context.[3] This framework allows us to gain a deep understanding of the economic agents' decision making while potentially confronted with a contractual trap; additionally, through investigating the agent's response, the principal

[1] Researchers have documented experimental evidence that human beings inevitably make mistakes while choosing among multiple options, even if they are fully aware that some options are better than the others; see, e.g., [9].

[2] It is worth mentioning that based on our definition of the *trap-filtered solution with cognition*, the agent does not exert cognitive effort, only if she sees a justifiable contract. In contrast, in [10], the agent will not exert cognitive effort, only if the principal opens the agent's eyes. Note also that in his framework, there is common knowledge of the game and rationality. This implies that the equilibrium contract is always justifiable. Nevertheless, cognitive effort still occurs even though justifiability is guaranteed. Please see Section 3 for details.

[3] A unified methodology in games is examined in [13].

can better design their contractual terms based on the economic implications generated in this paper. In Section 4, we use a stylized employee-employer contracting example to demonstrate the similarities and differences of those solution concepts. Through this example, we observe that the principal is able to exploit the agent by offering a non-justifiable contract when the agent passively updates her unawareness, but such an exploitation becomes impossible when the agent is able to reason how the principal fares upon offering such a "too-good-to-be-true" contract. Further, if the agent may interpret the non-justifiable contract as the principal's mistake, this exploitation is more likely to occur when the contractual traps are less common. The ability of exerting cognitive effort allows the agent to escape from a potential contractual trap, and the agent exerts more cognitive effort when the trap is more likely to happen.

Since we introduce the problem of unawareness in the principal-agent model, our paper is related to the vast literature on the unawareness. [14–16] first discuss the unawareness issue formally, and [17] show that it is impossible to model the non-trivial unawareness by using the standard state space. Nevertheless, [1–3] circumvent this negative result. The shared feature of these papers is that what is missing in the agent's mind is not arbitrary points in the state space, but rather, a *whole dimension* of it. We apply this idea to our contracting problems, as in [4,11,18]. Our principal-agent framework extends the standard moral hazard model; see, e.g., [19–22]. Unlike the aforementioned work, we incorporate the agent's unawareness, which gives rise to the novel issue of whether the principal should propose an incomplete contract. Further, our paper is also related to the literature on contractual incompleteness: verifiability ([23,24]), signaling ([25–27]), explicit writing costs ([28,29]) and inadequate cognition ([10,30,31]).

The remainder of this paper is organized as follows. In Section 2, we introduce the principal-agent framework, and in Section 3, we propose a number of solution concepts and discuss the behaviors under those solution concepts. In Section 4, we demonstrate the implications of these solution concepts in an example. Section 5 concludes.

2. The Model

We consider a stylized principal-agent model at a fairly general level.

Strategies. Let S_P and S_A denote the sets of strategies of the principal (P) and the agent (A), respectively. To incorporate the possibility that each party may determine decisions in many dimensions, here, $S_P \equiv A_P^1 \times \ldots \times A_P^M$ and $S_A \equiv A_A^1 \times \ldots \times A_A^N$ with M, $N < \infty$. In the canonical employee compensation example, the employer (the principal) may determine the compensation scheme that comprises the fixed payment and the commission rate for the employee (the agent). The employer may further determine other actions, such as the employee's housing option or retirement benefit. These decisions directly affect the utilities of the employer and the employee and are included in S_P. On the employee's side, she may have discretion in determining how much effort to exert in completing the project or whether to receive some external training that improves her productivity. The set, S_A, includes these decisions of the employee.

We use $s_P \equiv (a_P^1, \ldots, a_P^M)$ and $s_A \equiv (a_A^1, \ldots, a_A^N)$ to denote the elements in the strategy sets of the principal and the agent, respectively. Further, let $S \equiv S_P \times S_A$ with $s \in S$. To avoid technical difficulties, we assume that the set of strategy profiles, S, is finite.[4] Given the strategy profiles, s_P and s_A, the principal and the agent obtain utilities, u_P and u_A, respectively, where $u_i : S \mapsto \mathbb{R}$, $i \in \{P, A\}$is a mapping from the entire strategy profiles to a real-valued utility.[5] If, eventually, the agent rejects the contract, they receive the reservation utilities, $\overline{u}_P$ and $\overline{u}_A$, that correspond to the utilities they obtain from their outside options.

Unawareness. In contrast with the standard principal-agent models, we assume that the agent may be unaware of all the relevant aspects she or the principal is entitled to choose. Along the line of the modeling technique initiated by [3], let $D_i \equiv \{A_i^1, A_i^2, \ldots\}$ denote the collection of all action sets of party i, and $D \equiv D_P \cup D_A$ denotes the collection of all action sets of both the principal and the agent. Let W_i ($W_i \subseteq D_i$) denote the set of action sets of i of which the agent is aware *before* contracting, where $i \in \{P, A\}$. Thus, $W \equiv W_P \cup W_A$ represents the collection of action sets of which the agent is aware. We assume that the principal is fully aware of the entire set of strategy profiles, S, and knows the agent's awareness (i.e., W). In this sense, the principal is omniscient: he has a clear picture of the economic environment, and he knows precisely the agent's preference and what the agent is aware of.[6]

Since the agent may be unaware, the principal can determine whether to inform the agent via the contract offers. This contract offer may serve as an *eye-opener* that broadens the agent's vision and allows the agent to get a better understanding of the entire picture. Obviously, the principal must indicate in the contract the corresponding actions that the agent is aware of (i.e., W); additionally, the principal might announce actions of which the agent is unaware. We use $V \equiv V_P \cup V_A$ to represent the collection of action sets that are mentioned in the contract of which the agent is unaware. The set, V, can be interpreted as the principal's strategic announcement to alleviate the agent's unawareness.

Contract. We can now formally define a contract offered by the principal. In the following, we use the notation, $\times X$, to denote the Cartesian product of all action sets in $X \subseteq D$, i.e., $\times X \equiv \Pi_{Y \in X} Y$.

A **contract** is a vector, $\psi(V) \in \times(W \cup V)$, where $V \subseteq D \setminus W$.[7] Note that $\psi(V)$ specifies all actions that the agent is aware of *after* observing the contract. Let $\psi(V) \equiv (\psi_P(V), \psi_A(V))$, where $\psi_i(V)$ is composed only of party i's actions. Following the literature that incorporates unawareness into the contracting framework, we assume that whenever the principal announces some actions of which the

[4] In this way, the games will be finite, as well. The existence of a solution (under the appropriate solution concepts) can be easily established following the classical game theory literature.

[5] Notably, the general formulation here allows for uncertainty of outcomes. Whatever uncertainty there is regarding the outcome of a contract is assumed to be incorporated into the utility function, the utility should be viewed as the expected utility, with respect to a probability (assumed to be commonly known) on the contingencies. Of course, the agent may in general be unaware of some relevant contingencies, but here, we assume that the agent's unawareness is solely on actions.

[6] It is possible to extend our analysis to the case in which the principal is only partially aware, following the approach developed in [12]. Since our focus is on the impact of the agent's sophistication on the optimal contract design, we exclude the possibility of the principal's unawareness. Moreover, the situation where the principal is uncertain about the agent's awareness and, therefore, screens the agent's awareness is studied by [11] in the single-task optimal contract setting and [32] in the multi-task linear contract setting.

[7] The order of the elements is based on the following rule: The action sets of the principal precede the action sets of the agent, and $\forall i$, A_i^k precedes A_i^l, if and only if $k < l$. For example, if $W \cup V = \{A_A^2, A_P^3, A_P^1\}$, then $\times(W \cup V) \equiv A_P^1 \times A_P^3 \times A_A^2$.

agent is unaware, the agent is able to understand the contract immediately and adjust her awareness to account for the additional aspects specified in the contract; see, e.g., [4,5,11].

We can now define the contract completeness based on the above notion:

Definition 1. *A contract, $\psi(V)$, is **incomplete** in party i's strategy if $W_i \cup V_i \neq D_i$, where $i \in \{P, A\}$.*

By definition, a contract is incomplete in party i's strategy if it does not specify the complete *welfare-relevant* actions that party i can select.[8] We say a contract, ψ, is *incomplete* if ψ is incomplete in either the principal's or the agent's strategy. Given a contract, $\psi(V)$, the agent's effective strategy, denoted by $s_A(V)$, is confined within $\times(W_A \cup V_A)$;; likewise, $s_P(V) \in \times(W_P \cup V_P)$ corresponds to a feasible strategy profile for the principal *from the agent's perspective*. In general, $s(V) \equiv (s_P(V), s_A(V))$ is an incomplete strategy profile, since it is composed of the actions only in the agent's mind. The larger the set V is, the more dimensions the vector $s(V)$ has.

Although an incomplete contract does not specify the complete utility-relevant actions/obligations, it provides clear instructions of actions in some dimensions ($W_i \cup V_i$ for party i). If the actions are observable and are written in the contract, they are perfectly enforceable. Moreover, only these actions are enforced. In legal language, this corresponds to the extreme legal environment in which there are no mandatory or default rules in any dimension of parties' actions. The role of the court is passive in that it treats a written contract as complete and, thus, forbids all extrinsic evidence to clarify the ambiguity in the contract on the unspecified dimensions of actions.[9]

Rule-guided behavior. Since the contract is allowed to be incomplete, if $\psi(V)$ is incomplete in the agent's strategy, the agent can determine the actions specified in the contract accordingly, and she must "choose," *subconsciously*, the actions of which she is unaware. In this paper, we assume that if the agent is unaware of some aspect, $A_A^k \notin W_A \cup V_A$, after observing the contract, she subconsciously chooses her *default action* $\bar{a}_A^k$ in this aspect. Likewise, for $A_P^k \notin W_P \cup V_P$, the agent subconsciously assumes that the principal will choose the default action, $\bar{a}_P^k$. Since it is subconscious choice, it is natural to assume that the default action is unique. For example, if the agent may be unaware of playing "Second Life" in her office, the default action in this dimension is simply not playing it. If the agent may be unaware that the principal can delay the salary payments, the default action of the principal is not to delay them.

Let us elaborate more on the interpretation of the default actions. As the agent is unaware of A_A^k, the default action, $\bar{a}_A^k$, is chosen subconsciously based on her *rule-guided behavior* rather than her rational calculation. As in [34,35], the rule-guided behavior is orthogonal to the conscious process; the rule simply decodes the contractual situation facing the agent and gives an instruction, $\bar{a}_A^k$, to the agent. Since this rule is completely out of the agent's mind, the agent simply follows the rule without even noticing it. As an example, in the employee compensation problem, if an employee is unaware of the

[8] It is worth mentioning that [10] interprets contract completeness differently. Namely, he argues that a contract is more complete if the agent exerts more cognitive effort before contracting. In contrast, in our paper, a contract is incomplete if it does not specify all the utility-relevant actions.

[9] This coincides with the Willistonian or "textualist" approach, which argues that the contract is the only document that the court can use to determine the plain meaning of the contracting parties. In a nutshell, the court enforces only the letter, but not the spirit of the contract. See [33].

possibility of obtaining some training to improve her productivity, she may simply ignore the training without any contemplation. In such a scenario, receiving no training is her default action in this aspect.

The agent's unawareness is also reflected in how she perceives what the principal would do and how her own utility is affected. If the agent is unaware of A_P^k (i.e., $A_P^k \notin W_P \cup V_P$), the agent subconsciously takes for granted that the principal should choose $\bar{a}_P^k$ and, subconsciously, takes this default action, $\bar{a}_P^k$, into her own utility function. In this sense, the agent's conjecture of the principal's choice in the aspect she is unaware of is not based on rational expectation, but rather, on her *rule-guided perception.* This rule-guided perception can be regarded as a subconscious hypothesis in the agent's mind. The agent is unaware that this hypothesis could be wrong. In the example of the employee compensation problem, if the employee is unaware that her employer could remove the housing option, then the employee may contemplate whether to accept the contract as if the housing option is still inclusive. The employee's decision is based on this hypothesis, which may be wrong if *ex post* the employer does not provide the housing option.

In general, let us denote $s^C(V) \equiv (s_P^C(V), s_A^C(V)) \in \times(D \setminus (W \cup V))$ as the action profile that the agent is unaware of, where the superscript, C, stands for "complement". The complete strategy profile, $s = (s(V), s^C(V))$, is composed of both the strategy profiles in and out of the agent's mind. If the principal indeed chooses the default action in the aspects that the agent is unaware of, the strategy profile then satisfies that $s_P^C(V)$ consists of only default actions, $\bar{a}_P^k$. Define $\overline{s}(V) \equiv (\overline{s}_P(V), \overline{s}_A(V))$ as this special case. Note that the principal has the discretion to choose any feasible action in the aspect of which the agent is unaware. Thus, the principal's effective strategy space expands to the entire S_P. For example, if the obligation of an employer in the contract is only to fulfill the compensation level, then nothing prevents the employer from offering a low retirement benefit or postponing the salary payment.

Perceived utilities. Given the aforementioned, the agent's unawareness and rule-guided behavior (and perception), we can then articulate how the agent evaluates a contract, $\psi(V)$. Let $u_i^V : \times(W \cup V) \mapsto \mathbb{R}$, $i \in \{P, A\}$, denote the perceived utility function of party i *from the agent's viewpoint.* From the representation, the function, u_i^V, clearly depends on the strategy space, V, specified in the contract (and the corresponding actions, $s(V)$). In the presence of the agent's unawareness, we assume that:

$$u_i^V(\cdot) \equiv u_i(\cdot, \overline{s}(V)), i \in \{P, A\},$$

where $u_i : S \mapsto \mathbb{R}$ is the *actual* utility function of party i defined before. This reflects that the perceived utility functions, $u_i^V(\cdot)$, are coherent with the actual utility functions, $u_i(\cdot)$, where the missing variables are completed by the default strategy profile, $\overline{s}(V)$. Thus, the agent simply believes that the default actions will be taken in the aspects she is unaware of and derives the corresponding (perceived) utilities for herself and the principal.[10] Since, in our context, the agent updates/expands the perceived utility

[10] An alternative way to model the set of strategy profiles is to define a correspondence, $M : 2^S \mapsto 2^S$, from an announced subset of S, denoted by Y, to the updated action sets, $M(Y)$, in the agent's mind after the principal's announcement. Note that $M(Y) = M_P(Y) \times M_A(Y)$ specifies both the principal's and the agent's strategy sets. By this formulation, a contract, $\varphi = (\varphi_P, \varphi_A)$, is an element in $M(Y)$.

This alternative model sounds more general and flexible. However, it is not convenient to model how the principal deviates from his specified actions, φ_P, in the contract in a natural way. In fact, this deviation plays an important role in function to her actual utility, our notion follows the modeling strategy dating back to [36], where they study a general equilibrium framework.

As is standard in the principal-agent literature, we assume that the agent's actions are not observable, whereas all actions of the principal are verifiable.[11] Furthermore, we assume that the principal always intends to have the agent accept the contract as opposed to opting for his outside option. A sufficient, but crude, condition is that $\inf\limits_{s\in S} u_P(s) \geq \overline{u}_P$, where $\overline{u}_P$ corresponds to the principal's reservation utility as aforementioned. On the other hand, the agent may be better off turning down the contract offer. Specifically, if we define $\inf\limits_{s\in S} u_A(s)$ as the agent's worst-case utility level if she accepts the contract, this implies that $\inf\limits_{s\in S} u_A(s) < \overline{u}_A$. This assumption is adopted in the remainder of this paper. As we demonstrate later, this assumption simply rules out the trivial case in which the agent always accepts the contract, even if the principal may deceive her.

3. Solution Concepts

In order to predict the behaviors of the principal and the agent, it is essential to define what decision rules the principal and the agent should follow. In the terminology of game theory, these rules are described by "*solution concepts*". In the standard moral hazard model, in which every aspect is known to both parties, we can conveniently adopt subgame perfect Nash equilibrium as the solution concept. Since the game involves the agent's unawareness, subgame perfect Nash equilibrium is no longer appropriate. In the following, we first provide a preliminary discussion of the essential components and, then, introduce a number of solution concepts that are suitable for the economic environments that involve unawareness.

3.1. Preliminaries

Before introducing the solution concepts, we specify the timing in this contractual relationship as follows: (1) The principal proposes the contract, $\psi(V)$; upon observing the contract, the agent updates her awareness. (2) The agent decides whether to accept the contract. If not, the game is over, and both parties receive their reservation utilities from the outside options. (3) If the contract is accepted, the agent chooses s_A and subconsciously implements $\overline{s}_A(V)$ in $s_A^C(V)$; the principal chooses s_P afterwards.

We now introduce some definitions regarding a contract offer. Since there might be some discrepancy between the principal's suggested actions and the realized actions, we define $(\psi(V), s)$ as a *bundle*. Given the contract and the agent's updated awareness, we can describe how the agent chooses her strategy

the problem of contractual traps. On the contrary, our modeling framework avoids this difficulty, since the principal can freely choose any actions in the dimensions of which the agent is unaware, whereas the principal has to fulfill the actions in the dimensions in the contract of which the agent is aware.

[11] This may not be appropriate in certain scenarios, but modifications are straightforward. For example, if all actions are verifiable, the strategy of the agent can be directly written into the contract and enforceable. Since the principal can directly control the agent's actions, whether the agent is aware or not does not matter.

and whether to accept the contract or not. As in the standard principal-agent problems, the choice of the agent in the contract must be *incentive compatible* (IC):

$$\psi_A(V) \in \arg\max_{\widetilde{\psi}_A(V)} u_A^V(\psi_P(V), \widetilde{\psi}_A(V)), \tag{IC}$$

where $u_A^V(\psi_P(V), \widetilde{\psi}_A(V))$ on the right-hand side is the agent's perceived utility of a specific strategy profile, $\widetilde{\psi}_A(V)$ and $\psi_P(V)$ in the agent's mind. The incentive compatibility constraint guarantees that the strategy in the contract maximizes the agent's (perceived) utility.

Furthermore, in order to induce the agent to accept the contract, the following *individual rationality* (IR) constraint should hold:[12]

$$u_A^V(\psi(V)) \geq \overline{u}_A. \tag{IR}$$

We can now define the set of feasible contracts.

Definition 2. *A contract, $\psi(V)$, is **feasible** if it satisfies IC and IR. A bundle, $(\psi(V), s)$, is **coherent** if $\psi(V) = s(V)$ and $s_A^C(V) = \overline{s}_A(V)$.*

The coherence of a bundle ensures that the principal's realized actions are the same as his proposed actions in the contract and the agent chooses the default actions in the dimensions she is not aware of. Feasibility and coherence are required by all solution concepts in this paper.

3.2. Subgame-Perfect Solution

The first solution concept is the subgame-perfect solution, which essentially follows from the solution concept in the standard principal-agent problem.

Definition 3. *A bundle, $(\psi^*(V^*), s^*)$, is a **subgame-perfect solution** if the principal chooses V^*, $\psi^*(V^*)$ and s^* that maximize u_P s.t.$\psi(V)$ is feasible and $(\psi(V), s)$ is coherent.*

To interpret the subgame-perfect solution, it is helpful to first review the procedure to obtain the solution to a standard principal-agent problem. Without the issue of unawareness, this is done in two steps. In the first step, the contract must satisfy the incentive compatibility and individual rationality constraints (or collectively, the feasibility). In the second step, among the set of feasible contracts, the principal must select the one that maximizes his (expected) utility. When the agent is unaware of some aspects, there is room for the principal to determine what to announce/include in the contract. Thus the information conveyed in the contract must be optimal from the principal's perspective. Note that since we have assumed that $\inf\limits_{s\in S} u_P(s) \geq \overline{u}_P$, the principal strictly prefers to have the agent participate. Notably, as A_i^k is finite for all i and k, the game is finite as well. Thus, it is straightforward to obtain the following theorem.

[12] As a convention in the principal-agent models, if the IR constraint is satisfied, the agent is assumed to accept the contract [37].

Theorem 1. *There exists a subgame-perfect solution.*

The subgame-perfect solution can be regarded as a direct extension of the classical subgame perfect Nash equilibrium to incorporate the agent's unawareness. Recall that in a subgame perfect Nash equilibrium, at each node of the game, a player simply ignores how she reaches the node. All what matters is the future. Due to this subgame perfect feature, the game is solved by *backward induction*. Here, the novel feature is the agent's unawareness. Thus, the agent in our model must update her unawareness based on the principal's contract offer. The principal perfectly foresees the agent's response and, then, optimally determines the contract offer (and which set of actions to include in the contract).

However, because the principal and the agent perceive different games (due to the agent's unawareness), the agent may find the contract offer as suboptimal for the principal, even on the equilibrium path. This is in strict contrast with the standard game theory that assumes the common knowledge of the game. This discrepancy creates room for various choices of alternative solution concepts, as we elaborate in the subsequent sections.

3.3. Justifiable Solution

In the subgame-perfect solution, a critical assumption is that the agent takes the contract offered by the principal without thinking about whether the contract is indeed optimal for the principal. This does not cause any problem if the agent were fully aware of all the aspects. Nevertheless, as assumed in [4,5], an unaware agent may be reluctant to accept a contract if she believes that this contract is not the best contract (from the agent's viewpoint) among all the feasible contracts. This gives rise to the next solution concept, namely, the justifiable solution. The idea of this solution concept is first introduced in a framework of unforeseen contingencies by [4,5]. Before introducing the solution concept, let us first define a justifiable contract.

Definition 4. *A contract, $\psi(V)$, is **justifiable** if:*

- *it is feasible;*
- $\forall \widetilde{V} \subseteq V$, $\forall \psi(\widetilde{V}) \in \times(W \cup \widetilde{V})$, $\forall \widetilde{s}(V) \in \times(W \cup V)$*, such that* $\psi(\widetilde{V})$ *is feasible and* $(\psi(\widetilde{V}), \widetilde{s})$ *is coherent, we have* $u_P^V(\psi(V)) \geq u_P^V(\widetilde{s}(V))$.

According to the above definition, a contract is justifiable if the agent thinks that the principal indeed proposes an optimal contract. Note that since this can only be verified after the agent considers every possible contract that the principal would propose, an implicit assumption is that the agent is aware that something may go wrong.[13] Moreover, from the definition of a justifiable contract, the agent takes into consideration her own best response for every given contract. Thus, she believes that the principal can perfectly predict how she would behave (in the sense of subgame-perfect solution). This requires higher-order reasoning on the part of the agent. Notably, since the agent has only limited awareness, her calculation regarding the principal's utility is based on the agent's view of the principal's utility (u_P^V) rather than the principal's actual utility (u_P) and may be incorrect.

[13] Alternative models of awareness of unawareness are provided by [10,38,39].

When the agent is able to think that the principal indeed offers his optimal contract, her participation decision critically depends on whether the principal's contract offer is "reasonable". If based on the agent's investigation, the principal should have offered an alternative contract, the agent then suspects that something has gone wrong and, therefore, feels deceived, due to her unawareness. In such a scenario, whether the agent should accept the contract or not is determined by what utility she attaches to the contract. As the principal offers a contract that is not reasonable, an extremely pessimistic agent may assume the *worst case* scenario upon accepting the contract, which gives rise to the lowest utility, $\inf\limits_{s\in S} u_A(s)$. Since we assume that $\inf\limits_{s\in S} u_A(s) < \overline{u}_A$, the agent should reject the contract. Of course, the agent might not obtain $\inf\limits_{s\in S} u_A(s)$ when the contract is indeed a trap. However, since the agent does not know what the trap is—or, at least, the agent is unable to predict how the principal would behave given a "unreasonable" contract offer—it is convenient to assume that in the agent's mind, a contractual trap leads to the worst-case utility, $\inf\limits_{s\in S} u_A(s)$. The assumption that the agent knows her worst utility even if she is unaware can be justified by the limited liability of the agent.[14] Thus, we explicitly assume both pessimism and limited liability here.

One may argue that adopting the lowest utility level here is a special case of pessimism. In fact, as long as we assume that the agent's perceived utility, Z_A, from a non-justifiable contract is worse than her outside option, the agent will reject the contract anyway. Thus, the crucial assumption we make here is that an agent is pessimistic about a non-justifiable contract. Alternatively, one may consider the possibility that the agent can derive the worst possible outcome within her awareness, i.e., $Z_A = \inf\limits_{s_A\in W_A\cup V_A} u_A^V(s(V))$. This alternative scenario essentially makes no difference if the derived worst outcome within the agent's awareness is also worse than her outside option.[15]

The modified sequence of events is as follows. (1) The principal proposes the contract, $\psi(V)$; (2) the agent evaluates whether the contract is indeed the best interest of the principal; if not, she rejects the contract immediately; (3) after the agent's evaluation, if the contract is also optimal for the principal, the agent decides whether to accept the contract; (4) if the contract is rejected, both parties obtain their outside options; if it is accepted, the agent chooses s_A and subconsciously implements $\overline{s}_A(V)$ in $s_A^C(V)$; the principal then chooses s_P.

We next define the justifiable solution.

Definition 5. *A bundle,* $(\psi^*(V^*), s^*)$*, is a* ***justifiable solution*** *if the principal chooses* V^**,* $\psi^*(V^*)$ *and* s^* *that maximize* u_P *s.t.* $\psi(V)$ *is justifiable and* $(\psi(V), s)$ *is coherent.*

In a justifiable solution, we impose, on top of standard individual rationality and incentive compatibility constraints, the justifiability constraint on the principal's side. As the key difference between the subgame-perfect and justifiable solutions, this justifiability ensures that the principal offers the contract that is optimal for him based on the agent's calculation, and it significantly restricts the

[14] For example, for an investor, the worst outcome is zero return, which is usually known by the investor.

[15] Note that this alternative scenario has its own issue. Facing a non-justifiable contract, the agent is aware that something may go wrong and, therefore, she knows that her awareness is limited. Thus, it is no longer plausible that the agent still employs the derived worst outcome and uses this to compare with her outside option.

principal's choice of contract in order to induce the agent's participation. The existence of a justifiable solution can be easily established.[16]

Theorem 2. *There exists a justifiable solution.*

Proof. To prove the existence, first observe that there exists at least one justifiable contract: the one that makes the agent fully aware, although it may be suboptimal. Now, if the principal chooses the optimal contract among those feasible contracts that are justifiable for him, as well, we then obtain a justifiable solution according to the definition.

The idea of justifiable solution is similar to that of *forward induction* in game theory, as the player who currently plays also reasons about the motivation of the player who has just played upon observing the latter's actions. Recall that forward induction requires each player to rationalize other players' behaviors and actively interpret the rationale for an unintended action ([7,8]). In our context, since the principal is omniscient, but the agent is not fully aware, the idea of forward induction applies naturally to the agent, rather than the principal. The agent's reasoning upon receiving a contract alters what the principal is able to offer. Moreover, this solution concept is extremely restrictive in that any contract is rejected by the agent, as long as it does not qualify as justifiable. The agent's unwillingness to accept a non-justifiable contract follows from our assumption that $\inf_{s \in S} u_A(s) < \overline{u}_A$.[17]

It is worth mentioning that, in the agent's mind, the principal believes that the agent simply gives a best response to the contract (within the agent's awareness). In other words, the agent believes that the principal does not know that the agent can evaluate the justifiability of the contract. See also [5] for a discussion of justifiability in a different context. Our notion of justifiability is in the same spirit of the cognitive hierarchy (the generalized level-k thinking) discussed in [40]. More specifically, in our subgame-perfect solution, as the agent decides whether to accept the contract or not only based on her direct evaluation, her behavior can be interpreted as the optimal response of level-1 thinking. Anticipating the agent's response, the principal then designs the contract accordingly; this is in line with level-2 thinking. In the justifiable solution, the agent investigates whether the contract offered by the principal is indeed optimal from the principal's perspective, which in spirit coincides with level-3 thinking. Following this argument, we can then define the principal's corresponding response of optimal contract design as level-4. It is possible to extend our analysis to higher-level thinking, but such an extension is not realistic, as suggested in [40], and also complicates the presentation of the solution concepts.

So far, we have introduced two different solution concepts. In a subgame-perfect solution, the agent takes the contract as given and updates her awareness passively. In a justifiable solution, the agent rejects the contract whenever she thinks the principal does not offer the contract that is in the principal's best interest. These two solution concepts represent two extreme reactions from the agent's side in

[16] Note that in establishing Theorem 2, we adopt a totally different proof technique from [5].

[17] If this assumption is violated, i.e., $\inf_{s \in S} u_A(s) > \overline{u}_A$, the subgame-perfect solution suffices to be the appropriate solution concept, even if the agent is more sophisticated. In this sense, the forward induction step becomes unnecessary. The assumption that the agent knows the worst case is reasonable in several situations, such as the zero return case in the aforementioned investment example.

reasoning the principal's incentive. A natural question is whether there exist other solution concepts that lie between the two extremes. This motivates our next solution concept.

3.4. Trap-Filtered Solution

In a justifiable solution, we assume that as long as the agent finds that the contract is not justifiable, she believes that the principal is setting up a trap to take advantage of her, thereby rejecting the contract immediately. In this sense, from the agent's perspective, the principal is fully unreliable; on the other hand, the agent completely trusts the principal's rationality. This may appear to be a strong assumption in some scenarios. For example, it is possible that the agent believes that this contract simply results from *the principal's mistake*. Researchers have documented experimental evidence that human beings inevitably make mistakes while choosing among multiple options even if they are fully aware that some options are better than others. A growing stream of literature relates this to the "future uncertainty" and uses this to explain the "trembling-hand" behavior widely observed in the experiments. For example, this is done in *quantal response equilibrium* in [9].

Our goal in this section is to incorporate this type of bounded rationality into the unawareness framework. Formally, when the agent faces a contract, $\psi(V)$, that is not justifiable, she simply believes that with probability, $1-\rho$ it results from the principal's mistake, and with probability, ρ, this contract is a trap set up by the principal. We illustrate the determination of ρ in detail in Section 3.4.1. below. With these probabilities, we can then express the agent's expected utility upon observing a non-justifiable contract, $\psi(V)$, as follows:

$$U_A^T(\psi(V)) \equiv \rho Z_A + (1-\rho) u_A^V(\psi(V)),$$

where Z_A corresponds to the utility that the agent believes that she will obtain if the contract is a trap and $u_A^V(\psi(V))$ is the agent's utility after she updates her awareness and chooses the optimal strategies accordingly.

As discussed before, this utility, Z_A, can be set at some exogenous utility level the agent believes she will obtain after observing a non-justifiable contract. As before, a convenient way to assign a value to Z_A is based on extreme pessimism, in which case $Z_A = \inf_{s\in S} u_A(s)$ corresponds to the agent's worst-case utility. While there are other good candidates for Z_A that are less pessimistic, to fix ideas, in the sequel, we choose $Z_A = \inf_{s\in S} u_A(s)$ for ease of exposition.

Given the agent's belief about the principal's behavior, the agent accepts the contract, $\psi(V)$, if the following individual rationality constraint is satisfied:

$$U_A^T(\psi(V)) \geq \overline{u}_A. \tag{IR-T}$$

We can now define an acceptable contract when the agent believes in the possibility of the principal's mistake and the corresponding solution concept.

Definition 6. *A contract, $\psi(V)$, is **trap-filtered** if 1) it is justifiable or 2) it is feasible and IR-Tis satisfied. A bundle, $(\psi^*(V^*), s^*)$, is a **trap-filtered solution** if the principal chooses V^*, $\psi^*(V^*)$ and s^* that maximize u_P s.t. $\psi(V)$ is trap-filtered, and $(\psi(V), s)$ is coherent.*

The idea behind the trap-filtered contract is that the agent believes that the principal may cheat her only if the contract is not justifiable. In such a scenario, a non-justifiable contract makes the agent suspect whether it is indeed in the best interest of the principal, thereby giving rise to the second set of condition. Note that neither condition is implied by the other: It is possible that a justifiable contract does not satisfy IR-T, and a contract that satisfies condition IR-T need not be justifiable.

In Section 3.4.1., we provide an interpretation of this solution concept through a game with a lexicographic probabilistic system by [41]. We can extend the proof of the existence of a justifiable solution, in a straightforward manner, to guarantee the existence of a trap-filtered solution.

Theorem 3. *There exists a trap-filtered solution.*

Note that when $\rho = 0$, the agent is extremely confident that any non-justifiable contract should be attributed to the principal's mistake; she proceeds to update her awareness according to the contract and determines her optimal strategies, and the trap-filtered solution degenerates to a subgame-perfect solution. On the other hand, if $\rho = 1$, the agent believes that the principal never makes a mistake; thus, whenever she sees a non-justifiable contract, she perceives it as a trap and the trap-filtered solution coincides with a justifiable solution. Thus, the trap-filtered solution can be regarded as a broader family of the solution concepts that incorporate the ones reported in the literature.

3.4.1. Interpretation of the Trap-Filtered Solution

One may wonder why the agent assigns a positive probability ρ to the event where this contract is a trap only upon receiving a non-justifiable contract. We now provide a micro-foundation to interpret the trap-filtered solution.[18]

Consider the following scenario. Suppose that from the agent's viewpoint, there are two types of principals: a normal one (with probability one) and a "crazy" one (with probability zero). The crazy principal always makes a mistake (i.e., offering a non-justifiable contract). However, a normal principal who is rational may intentionally set up a trap for the agent. There are also two types of games: the one where the agent knows the game (with probability one) and the one the agent is unaware of something (and, thus, does not know the actual game) (with probability zero). The agent is uncertain of the principal's type and her knowledge of the game, and the values of these two variables are independent. Here, we use the *lexicographic probabilistic system* by [41]. First, the event that the principal is normal is first-order, and the event of having a crazy principal is second-order. Second, the event that the agent knows the game is first-order, and the event that the agent is unaware of the game is second-order. Further, the event where the principal is crazy and the event in which the agent is unaware of something are of the same order.

Formally, the state space consists of four states, ω_1, ω_2, ω_3 and ω_4, which are differentiated by whether the principal is normal or crazy and whether the agent is fully aware of the entire game. Specifically, let ω_1 represent the state in which the principal is normal and the agent knows the entire game; ω_2 represents the state in which the principal is normal and the agent is unaware of something; ω_3 represents the state in

[18] The interpretation automatically applies to the trap-filtered solution with cognition described below, as it can be seen as a special case in which cognitive effort is infinitely costly.

which the principal is crazy and the agent knows the entire game; finally, ω_4 represents the state in which the principal is crazy and the agent is unaware of something. Given the four states, the lexicographic probabilistic system, $\mu = (p_1, p_2)$, is as follows: $p_1(\omega_1) = 1$; $p_1(\omega_2) = p_1(\omega_3) = 0$; and $p_2(\omega_2) = \rho$; $p_2(\omega_3) = 1 - \rho$. Put differently, in the terminology of [41], we have assumed that $\omega_1 >_\mu \omega_2$ and $\omega_1 >_\mu \omega_3$.

Thus, if the agent faces a justifiable contract, she believes that with probability one, she knows the game and the principal is normal. However, conditional on a non-justifiable contract, the agent believes that there is a trap with probability ρ, and it results from the principal's mistake with probability $1 - \rho$. Therefore, the agent's optimal behavior is exactly as described in the trap-filtered solution.[19]

3.5. Trap-Filtered Solution with Cognition

The trap-filtered solution has nicely synthesized all possible scenarios regarding how the agent perceives the principal's contract offer. Nevertheless, in all the aforementioned solution concepts, the agent can only passively interpret the principal's behavior and react accordingly, based on her conservativeness and confidence. While this might be satisfactory in certain scenarios, it could also be possible that the agent is able to "think" through the scenarios upon receiving a contract. Of course, if the contract offer is justifiable, such *cognitive effort* does not benefit the agent, since there is no trap with a probability of one, as we discussed above in Section 3.4.1.; however, if the principal indeed offers a non-justifiable contract, thinking allows the agent to pull back from being trapped into a contract. As in [10], such cognitive effort is typically costly, and the associated cost is implicit and frequently ignored in classical contract theory. Our goal in this section is to incorporate the cognition into our contractual framework with unawareness.

To formalize our ideas, we assume that the agent can afford to spend some resources to evaluate whether a non-justifiable contract is due to the principal's mistake or the agent's lack of awareness. This cognition stage arises after the principal has offered the contract, but before the agent decides whether to accept the contract. The more the agent spends (after the contract is announced), the more likely she is able to identify a contractual trap given that there is indeed a trap. Specifically, let $c \in [0, 1]$ denote the probability that the agent finds out that the contract is a trap (conditional on the event that it is indeed a trap). The associated cost of cognitive effort is denoted by an increasing function, $T(c)$.[20] Note that even though the agent actively thinks through the scenarios, it is still possible that the principal may trap the agent via a non-justifiable contract (but less likely, due to the agent's cognitive effort).[21]

[19] An alternative interpretation is that the "crazy" principal exists with a positive probability, but the agent is initially unaware of this event together with the event that he is unaware of actions. Upon observing a non-justifiable contract, however, the agent becomes aware of both events. While this interpretation is natural, it entails the agent's extra unawareness of contingencies. Since we focus on the agent's unawareness of actions in this paper, we tend to minimize the agent's bounded rationality in other aspects. Thus, we prefer to adopt the interpretation with the lexicographic probabilistic system.

[20] Note that it is innocuous to claim to incorporate cognition here. We abstract away from the detailed process of cognition. Yet, the economically relevant effect of cognition is modeled by the reduced form of the cognition cost function.

[21] Note that in the cognition stage, we have implicitly assumed that it is costless to determine whether or not a contract is justifiable. In this sense, the cost of information processing in this regard is omitted. In principle, it is possible

With the addition of the cognition stage, the modified sequence of events is as follows. (1) The principal proposes a contract, $\psi(V)$. (2) Upon receiving the contract, the agent (costlessly) evaluates whether the contract is justifiable. (3) If the contract is justifiable, the agent spends no cognitive cost and determines directly whether to accept the contract; if the contract is non-justifiable, the agent makes the cognitive effort to evaluate whether the contract is a trap or simply a principal's mistake. (4) After the cognition stage, if the agent figures out that a non-justifiable contract is a trap, she refuses to sign the contract, and the game ends immediately; if, based on her cognitive effort, the agent thinks it is more likely to be the principal's mistake, she then determines whether to accept the contract.[22] (5) Finally, if the contract is accepted, the principal and the agent make their decisions and obtain their utilities.

Let us articulate how the agent decides whether to accept the contract. Suppose that the agent decides to bear the cognitive cost, $T(c)$, upon observing a non-justifiable contract. Then, with probability $1-\rho$, the agent believes that this comes entirely from the principal's mistake. In this case, she obtains utility $u_A^V(\psi(V))$ upon accepting the contract. With probability ρc, the agent figures out that the contract is an intentional trap. To be consistent with the scenarios discussed earlier, we assume that the agent rejects the contract if she thinks it is a trap and obtains her reservation utility, $\overline{u}_A$. Finally, with probability $\rho(1-c)$, the agent cannot figure out the trap and attaches *ex ante* the utility, $\inf\limits_{s\in S} u_A(s)$, to such an event. The agent's pre-cognition expected utility is:

$$U_A^C(c,\psi(V)) \equiv \rho c\overline{u}_A + \rho(1-c)\inf_{s\in S} u_A(s) + (1-\rho)u_A^V(\psi(V)) - T(c).$$

This determines the optimal cognitive effort spending as follows:

$$c^*(\psi(V)) \in \arg\max_{c\in[0,1]} U_A^C(c,\psi(V)),$$

and the corresponding optimal expected utility is:

$$U_A^C(\psi(V)) \equiv U_A^C(c^*(\psi(V)),\psi(V)).$$

The agent will accept a non-justifiable contract, $\psi(V)$, if and only if the following pre-cognition individual rationality constraint holds:

$$U_A^C(\psi(V)) \geq \overline{u}_A. \qquad \text{(IR-C)}$$

Note also that cognitive effort allows the agent to determine whether to accept the contract *after* the cognition stage. We can also write down the post-cognition individual rationality constraint (after the cognition stage). It requires that:

$$\frac{\rho(1-c)}{\rho(1-c)+1-\rho}\inf_{s\in S} u_A(s) + \frac{1-\rho}{\rho(1-c)+1-\rho}u_A^V(\psi(V)) \geq \overline{u}_A. \qquad \text{(IR-C2)}$$

that evaluating the details of the contract to verify its justifiability also requires the agent to spend cognitive effort, as demonstrated in [42]. However, determining whether to evaluate each step itself may then require the agent to spend her own cognitive effort, and determining whether to determine to evaluate each step may again require cognitive effort. This leads to endless iterations and distracts attention from the main issues of this paper.

[22] This is different from [10] in that cognitive effort occurs when the contract is not justifiable, whereas in [10], the agent exerts cognitive effort only when the agent's eyes are not opened. In essence, a non-justifiable contract and non-eye-opening information play the same role in the situations where something may go wrong for the agent. However, our formulation allows the possibility of seeing a "too-good-to-be-true" contract that would never occur in [10].

In IR-C2, $\rho(1-c)+1-\rho$ is the probability that the agent does not find any evidence of the contractual trap, and $\frac{\rho(1-c)}{\rho(1-c)+1-\rho}$ and $\frac{1-\rho}{\rho(1-c)+1-\rho}$ are the conditional probabilities that a non-justifiable contract is indeed a trap or a result of the principal's mistake, respectively.

Because IR-C implicitly assumes that the agent will accept the contract *ex post*, it is straightforward to obtain the following theorem.

Theorem 4. *IR-C implies IR-C2.*

We can now introduce the solution concept with cognition.

Definition 7. *A contract, $\psi(V)$, is **trap-filtered with cognition** if it is justifiable or it is feasible and IR-C holds. A bundle, $(\psi^*(V^*), s^*, c^*)$, is a **trap-filtered solution with cognition** if the principal chooses V^*, $\psi^*(V^*)$ and s^* that maximize $\{c^*\overline{u}_p + (1-c^*)u_P(s)\}$ s.t. $\psi(V)$ is trap-filtered with cognition, $(\psi(V), s)$ is coherent, $c^* = 0$ if $\psi(V)$ is justifiable and $c^* \in \arg\max_{c' \in [0,1]} U_A^C(c', \psi(V))$, otherwise.*

Along the lines of Section 3.4.1. , there are two types of principals: a normal one and a "crazy" one, and a normal principal may intentionally set up a trap for the agent. Thus, in the definition of the trap-filtered solution with cognition, the normal principal intends to choose $V^*, \psi^*(V^*)$ and s^* that maximize:

$$c^*\overline{u}_p + (1-c^*)u_P(s),$$

where the term, $c^*\overline{u}_p$, corresponds to the case in which the cognitive effort is effective (which occurs with probability c^*), and the second term corresponds to the case in which the agent accepts the contract and makes the optimal actions accordingly. In response to the potential contractual trap from the normal principal, the agent exerts the optimal cognitive effort to figure out whether there is a contractual trap. Upon receiving a non-justifiable contract, in the agent's mind, there is a distinction between two cases: (1) the principal makes a mistake (which occurs with probability $1-\rho$); and (2) the principal indeed sets up a trap, but the agent fails to catch it (with probability $\rho(1-c)$).

Note that in this formulation, the cognitive effort can take an arbitrary value in the interval, $[0,1]$. This implies that the game is no longer finite. Nevertheless, a finite game is sufficient to show that a trap-filtered solution with cognition exists. As we demonstrate in Section 4, a trap-filtered solution with cognition may still exist, even if the strategy space is not finite. In general, if T is weakly convex and continuous, we can follow the arguments in [43–45] to establish the existence of a trap-filtered solution with cognition. We highlight this result below.

Theorem 5. *There exists a trap-filtered solution with cognition.*

3.6. Payoff Ordering for the Principal

We have introduced a sequence of solution concepts that assume different degrees of rationality and cognitive ability on the agent. We now check if the principal's payoff can be ordered according to the solution concepts. It is clear that the subgame-perfect solution gives the principal the highest payoff, because it requires only the feasibility of the contract; thus, it imposes the weakest constraint for the principal's problem. In contrast, the justifiable solution generates the lowest payoff of the principal, since justifiability of the contract is necessary, whereas it is only sufficient in all other solution concepts.

Thus, the principal's payoffs in the other two solution concepts are in between. Notice that the principal's payoff in the trap-filtered solution with cognition must be (weakly) lower than that without cognition. This is because one can interpret the trap-filtered solution as an extreme case of that with cognition when cognition is infinitely costly. Thus, the trap-filtered solution with cognition entails a (weakly) higher probability of rejection of the contract for each contract, thereby hurting the principal's profit. This gives rise to the following theorem.

Theorem 6. *The principal's payoff ordering according to the solution concepts is: subgame-perfect solution $\succ$ trap-filtered solution $\succ$ trap-filtered solution with cognition $\succ$ justifiable solution.*

4. Contractual Traps: A Numerical Example

In this section, we demonstrate the differences among these solution concepts via a numerical example.

Problem descriptions. In this example, both parties have two dimensions of strategies. The first dimension action set, A_i^1, is a singleton, $\{\overline{a}_i^1\}$, that consists of a usual action of the party i. The agent is, however, unaware of the second dimension of actions. For simplicity, let us assume that $A_A^2 = \{0, 1\}$ and $A_P^2 = \{0, 2\}$, and the default actions are $\overline{a}_P^2 = \overline{a}_A^2 = 0$. The alternative actions, $a_P^2 = 2$ and $a_A^2 = 1$, are the unforeseen actions for the agent. In our notation, $W = \{A_P^1, A_A^1\}$, since the agent is only aware of the usual actions of both parties in the first dimension.

To visualize this example, suppose that a principal (employer) intends to hire an agent (potential employee). We can interpret the first dimension as the typical reception from the principal as he negotiates with the agent. In the second dimension, the agent's choice (if she is aware) is between a status quo contract and a novel contract, and the principal's corresponding action is whether to provide the on-campus faculty housing in the employment contract. In the seasoned faculty hire case (such as an advanced assistant professor's move), the status quo contract may be a tenure-track assistant professor with a typical startup package that includes the on-campus faculty housing option. The novel contract can be a specially-arranged untenured associate professor position. The agent's default action ($\overline{a}_A^2 = 0$) is to choose a status quo contract, and the principal's default action ($\overline{a}_P^2 = 0$) is to provide the faculty housing. Further, assume that the principal must provide the faculty housing in the status quo contract, but he is able to remove it from the novel contract.[23] The alternative action, $a_A^2 = 1$, corresponds to the case in which the agent chooses a novel contract, and $a_P^2 = 2$ corresponds to the principal's decision to remove the faculty housing option from the novel contract. This saves the principal's cost, but reduces the agent's utility upon joining.

Given the two dimensions of actions, the actual utilities of the principal and the agent are, respectively, $u_P = a_P^1 a_A^1 - a_A^2 + a_A^2 a_P^2$ and $u_A = a_P^1 a_A^1 + a_A^2 - a_A^2 a_P^2$. Since A_i^1 is a singleton, we can conveniently assume that the default (regular) actions are both one (i.e., $\overline{a}_P^1 = \overline{a}_A^1 = 1$). After these substitutions, we obtain that $u_P = 1 - a_A^2 + a_A^2 a_P^2$ and $u_A = 1 + a_A^2 - a_A^2 a_P^2$. Let the reservation utilities of them be

[23] Housing options are rather subtle in various schools in Hong Kong and Singapore, where the housing market is extremely expensive. Our assumption ensures that if the principal intends to set up a trap, he can only do so upon introducing the novel contract. If the principal is also allowed to remove the faculty housing secretly from a status quo contract, the trap could appear in all scenarios.

$\overline{u}_P = \delta$ and $\overline{u}_A = 1$ (which correspond to the situation in which no trade occurs). Note that in order to guarantee that the principal always intends to induce the agent's participation, we require that $\delta < 0$.

Subgame-perfect solution. Let us first consider the scenario in which the principal does not announce any new actions (the option of signing a novel contract) to the agent, i.e., $V = \emptyset$. In such a scenario, the agent can only decide between signing the status quo contract ($a_A^2 = 0$) and simply walking away. Given this, since the principal cannot remove the faculty housing option, the principal's action affects neither the agent nor the principal himself. Therefore, choosing $a_P^2 = 0$ is the principal's best response, and as a result, both the principal and the agent obtain a utility of one.

Now suppose that the principal informs the agent of the possibility of choosing the novel contract (i.e, $V = \{A_A^2\}$); the agent is then aware of this new dimension and, therefore, makes the decision optimally based on her perceived utility. In this case, since the principal does not disclose his own action set, A_P^2 (that he may remove the faculty housing option), under the subgame-perfect solution, the agent continues to (subconsciously) believe that the principal will provide the faculty housing option ($a_P^2 = 0$). Thus, from the agent's perspective, her perceived utility is $u_A^{A_A^2} = 1 + a_A^2$. The corresponding best response is to choose $a_A^2 = 1$, and in the agent's mind, she should obtain a perceived utility of two.

We now turn to the principal's problem. By backward induction, the principal perfectly foresees the agent's action, $a_A^2 = 1$. Consequently, his (actual) utility becomes $u_P = a_P^2$, and thus, his optimal strategy is to choose $a_P^2 = 2$. From the above discussion:

$$(\psi(V), s) = ((\overline{a}_P^1, \overline{a}_A^1, 1), (\overline{a}_P^1, 2, \overline{a}_A^1, 1))$$

is the unique subgame-perfect solution. *The principal proposes the novel contract for the agent, but does not mention the possibility of removing the faculty housing option.* Notably, this solution concept gives rise to a utility of two for the principal, but an actual utility $1 + a_A^2 - a_A^2 a_P^2 = 0$ for the agent, whereas in the agent's mind, the supposed utility is two rather than zero. In this sense, the contract, $\psi(V)$, with $V = \{A_A^2\}$ is a *contractual trap* for the agent. The agent takes the lure of the novel contract and, thus, is willing to choose $a_A^2 = 1$. The principal then takes advantage of the agent by removing the faculty housing option ($a_P^2 = 2$).

Justifiable solution. The discussion above demonstrates how a contractual trap can be implemented, even if the agent is fully naive (and suffers from her lack of awareness). We next apply the idea of justifiability to this example. When the agent is sophisticated, she may feel that the novel contract is "*too-good-to-be-true*". This is because, in the agent's mind, if A_A^2 were not specified in the contract, the principal would receive utility $u_P^{A_A^2} = 1$. However, the contract with $V = \{A_A^2\}$ offers the agent an opportunity to choose an action, a_A^2, which benefits the agent herself, but might hurt the principal as the principal receives utility $u_P^{A_A^2} = 0$. Thus the contract in the subgame-perfect solution is *not justifiable*. Note that from the agent's perspective, the principal's utility crucially depends on the offered contract. When $V = \emptyset$, the agent believes that $u_P = 1$; when $V = \{A_A^2\}$, it becomes $u_P = 1 - a_A^2$; when $V = \{A_P^2,\ A_A^2\}$, the perceived utility becomes $u_P = 1 - a_A^2 + a_A^2 a_P^2$. Note that $\inf_{s \in S} u_A(s) = 0 < 1 = \overline{u}_A$. Thus, the agent will indeed reject this non-justifiable contract.

Trap-filtered solution. Next, we assume that the agent believes that the non-justifiable contract (regarding the novel contract) may result from the principal's mistake (with probability $1 - \rho$). It follows from straightforward algebra that the contract with $V = \{A_A^2\}$ is a trap-filtered solution when $\rho \leq \frac{1}{2}$.

When the probability of the principal's mistake is high, upon receiving a non-justifiable contract, the agent is more inclined to interpret it as a mistake and, consequently, accepts the contract with $V = \{A_A^2\}$ *although it is too-good-to-be-true.* In other words, the principal sets a trap only when the agent believes that the non-justifiability of the contract is more likely due to the principal's mistake rather than a trap. This coincides with our intuition: in a society where contractual traps are not common, the agent is more inclined to accept non-justifiable contracts. On the principal's side, the normal principal is (weakly) better off for a higher ρ, as the trap is easier to implement, because the principal may receive a utility of two rather than one (the utility given a smaller ρ).

Trap-filtered solution with cognition. Finally, we introduce the cognitive effort. When confronted with two options to choose from (the existing and the new contract), if the agent goes through the cognition stage, she may be able to come up with effective ways to evaluate whether a trap is hidden in the contract. The outcome of the agent's cognitive effort in this case would eliminate the possibility of a trap, because it forces the principal to fully specify the contract for the new contract.

To demonstrate our idea, we let $T(c) = \frac{1}{2}c^2$ for simplicity. Following the definition of the trap-filtered solution with cognition, the agent accepts the contract only if:

$$\max_{c \geq 0} \left\{ \rho c + 2(1-\rho) - \frac{1}{2}c^2 \right\} \geq 1,$$

that is, $\rho \leq 2 - \sqrt{2} \approx 0.58579$, which is greater than the previous cutoff value, 0.5.[24] Since the agent's ability to exert cognitive effort allows her to reject a contract after the cognition stage, it is conceivable that she can more likely afford to accept the contract (i.e., with a higher ρ compared to the case without the cognitive effort), and the agent should obtain a higher expected utility. We further find $c = \rho$, i.e., the more likely there is a trap, the more effort the agent exerts in cognition.

In the presence of the cognition stage, the principal sets up a trap only if $2(1-c) + c\delta \geq 1$, that is, $\rho(2-\delta) \leq 1$. This also has an intuitive interpretation. When the principal's outside utility (δ) is low, he is severely punished by the agent's non-participation once the contractual trap is caught. Thus, the principal's incentive to set a contractual trap declines as δ becomes lower. As in the case without the cognition stage, we also observe that the possibility of setting a contractual trap is higher when the agent is more convinced that this results from the principal's mistake (ρ is low).

To summarize, if the agent is able to exert cognitive effort, $(\psi(V), s) = ((\overline{a}_P^1, \overline{a}_A^1, 1), (\overline{a}_P^1, 2, \overline{a}_A^1, 1))$ is a trap-filtered solution with cognition if $\rho(2-\delta) \leq 1$ and $\rho \leq 0.58579$. Note that in this example, the support of the cognitive effort is continuous rather than finite. However, a trap-filtered solution with cognition still exists.

In this example, we observe that the principal is able to exploit the agent by offering a non-justifiable contract when the agent passively updates her unawareness, but such an exploitation becomes impossible when the agent is able to reason how the principal fares upon offering such a "too-good-to-be-true" contract. Further, if the agent may interpret the non-justifiable contract as a principal's mistake, this

[24] Note that this result is quite general in the sense that it does not depend on the particular cost function, $T(c)$, we employed. Let $T(0) = 0$. Since the agent accepts the contract if $\max_{c \geq 0} \{\rho c + 2(1-\rho) - T(c)\} \geq 1$, i.e., $\rho c^* - T(c^*) \geq 2\rho - 1$ where c^* is the optimal level of cognition and $\rho c^* - T(c^*) \geq 0$ (otherwise, the agent chooses the zero cognition level), we have that the cutoff value of ρ being greater than 0.5.

exploitation is more likely to occur when the contractual traps are less common. The ability of cognition allows the agent to escape from a potential contractual trap sometimes, and the agent exerts more cognitive effort when the trap is more likely to happen.

5. Concluding Remarks

In this paper, we provide a general contracting framework to investigate strategic interactions in the presence of unawareness, reasoning and cognition and propose several solution concepts in various degrees of the agent's sophistication. These solution concepts are well suited in various economic contexts that involve the contracting parties' unawareness, bounded rationality, psychological effect and cognition. The primary message we intend to convey in this paper is to demonstrate the possibility of incorporating unawareness, reasoning and cognition in a single framework. This general framework certainly has its own limitations; however, due to its simplicity, we open up a number of possible extensions for other economic contexts of interest. For example, we abstract away from the renegotiation problem in the post-contracting stage. Nevertheless, if an agent figures out that the principal's contract is non-justifiable, it is conceivable that the two contracting parties may attempt to communicate and renegotiate the contract. In such a scenario, alternative solution concepts may be proposed, and it would be intriguing to see whether this renegotiation stage influences the agent's response to the contract offer and how the principal designs the optimal contract.

Our focus on the monopolistic principal's optimal contract design problem may be a bit excessive. In certain situations, it is possible that multiple principals, either homogeneous or heterogeneous in terms of their awareness and preferences, may compete in hiring an unaware agent. Thus, the agent's awareness in the post-contractual stage is jointly determined by the contracts offered by these principals with conflicting interests. Another possible extension is to introduce multiple agents with heterogeneous degrees/dimensions of unawareness. The interesting question in this alternative setting is whether the principal intends to offer secret/private contracts to these agents, and if so, whether the agents have an incentive to communicate with each other after receiving the principal's offers.

In this paper, we focus on the one-shot transaction between the principal and the agent. However, in many practical situations, these contracting parties may interact in multiple rounds. While extended to the multiple-round (repeated) setting, the optimal contract design in this principal-agent relationship becomes more sophisticated. It has been well-documented that in a dynamic contracting environment, the ratchet effect and the commitment problem significantly complicate the optimal contract design. In our framework, we impose, on top of those difficulties, the additional strategic concerns of how much information to disclose through the contract offers over time and how much information the agent is able to infer/reason/think about given the sequence of proposed contracts. Finally, given the principal's incentive to offer the contractual trap and the agent's (wasteful) effort on cognition, it might be welfare improving if a benevolent third party is introduced to control the information flow. While our framework certainly provides some preliminary exercise, thorough studies on social welfare, efficiency and fairness are needed in order to provide a general picture and are left for future research.

Acknowledgments

An extended abstract with the title, “Contractual Traps”, is included in the proceedings of the 12th Theoretical Aspects of Rationality and Knowledge (TARK XII). We thank the reviewers for many valuable comments and suggestions that significantly improved the quality of the paper. We also thank the participants of TARK XII at Stanford and the 20th International Conference on Game Theory at Stony Brook, Selina Cai, Chia-Hui Chen, Spyros Galanis, Xing Hu, Jing Li and, especially, Burkhard C. Schipper for very helpful comments. All remaining errors are our own.

References

1. Galanis, S. Unawareness of theorems. *Econ. Theory* **2013**, *52*, 41–73.
2. Heifetz, A.; Meier, M.; Schipper, B. Interactive unawareness. *J. Econ. Theory* **2006**, *130*, 78–94.
3. Li, J. Information structures with unawareness. *J. Econ. Theory* **2009**, *144*, 977–993.
4. Filiz-Ozbay, E. Incorporating unawareness into contract theory. *Games Econ. Behav.* **2012**, *76*, 181–194.
5. Ozbay, E. Unawareness and Strategic Announcements in Games with Uncertainty. Working paper, University of Maryland, 2008.
6. Heifetz, A.; Meier, M.; Schipper, B. Dynamic unawareness and rationalizable behavior. Working paper, The Open University of Israel, 2009.
7. Battigalli, P.; Siniscalchi, M. Strong belief and forward induction reasoning. *J. Econ. Theory* **2002**, *106*, 356–391.
8. Kohlberg, E.; Mertens, J.F. On the strategic stability of equilibria. *Econometrica* **1986**, *54*, 1003–1037.
9. McKelvey, R.; Palfrey, T. Quantal response equilibria for normal form games. *Games Econ. Behav.* **1995**, *10*, 6–38.
10. Tirole, J. Cognition and incomplete contracts. *Am. Econ. Rev.* **2009**, *99*, 265–294.
11. Von Thadden, E.; Zhao, X. Incentives for unaware agents. *Rev. Econ. Stud.* **2012**, *79*, 1151–1174.
12. Zhao, X. Moral hazard with unawareness. *Ration. Soc.* **2008**, *20*, 471–496.
13. Feinberg, Y. Games with Unawareness. Working paper, Stanford University, 2010.
14. Fagin, R.; Halpern, J. Belief, awareness, and limited reasoning. *Artif. Intell.* **1988**, *34*, 39–76.
15. Modica, S.; Rustichini, A. Awareness and partitional information structures. *Theory Decis.* **1994**, *37*, 107–124.
16. Modica, S.; Rustichini, A. Unawareness and partitional information structures. *Games Econ. Behav.* **1999**, *27*, 265–298.
17. Dekel, E.; Lipman, B.; Rustichini, A. Standard state-space models preclude unawareness. *Econometrica* **1998**, *66*, 159–174.
18. Gabaix, X.; Laibson, D. Shrouded attributes, consumer myopia, and information suppression in competitive markets. *Q. J. Econ.* **2006**, *121*, 505–540.
19. Grossman, S.; Hart, O. An analysis of the principal-agent problem. *Econometrica* **1983**, *51*, 7–45.
20. Holmstrom, B. Moral hazard and observability. *Bell J. Econ.* **1979**, *10*, 74–91.

21. Holmstrom, B.; Milgrom, P. Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *J. Law Econ. Organ.* **1991**, *7*, 24–52.
22. Mirrlees, J. The theory of moral hazard and unobservable behaviour: Part I. *Rev. Econ. Stud.* **1999**, *66*, 3–21.
23. Grossman, S.; Hart, O. The costs and benefits of ownership: A theory of vertical and lateral integration. *J. Polit. Econ.* **1986**, *94*, 691–719.
24. Hart, O.; Moore, J. Property rights and the nature of the firm. *J. Polit. Econ.* **1990**, *98*, 1119–1158.
25. Aghion, P.; Bolton, P. Contracts as a barrier to entry. *Am. Econ. Rev.* **1987**, *77*, 388–401.
26. Chung, K.; Fortnow, L. Loopholes. Working paper, University of Minnesota and Northwestern University, 2007.
27. Spier, K. Incomplete contracts and signalling. *RAND J. Econ.* **1992**, *23*, 432–443.
28. Anderlini, L.; Felli, L. Incomplete contracts and complexity costs. *Theory Decis.* **1999**, *46*, 23–50.
29. Dye, R. Costly contract contingencies. *Int. Econ. Rev.* **1985**, *26*, 233–250.
30. Zhao, X. Strategic Mis-selling and Pre-Contractual Cognition. mimeo, Hong Kong University of Science and Technology, 2012.
31. Bolton, P.; Faure-Grimaud, A. Satisficing contracts. *Rev. Econ. Stud.* **2010**, *77*, 937–971.
32. von Thadden, E.; Zhao, X. Multi-task Agency with Unawareness. mimeo, Hong Kong University of Science and Technology, 2013.
33. Schwartz, A.; Scott, R. Contract theory and the limits of contract law. *Yale Law J.* **2003**, *113*, 541–619.
34. Hayek, F.; *Rules, Perception and Intelligibility*; Studies in Philosophy, Politics and Economics, The University of Chicago Press: Chicago, IL, USA, 1967; pp. 43–65.
35. Vanberg, V. Rational choice *vs*. program-based behavior: Alternative theoretical approaches and their relevance for the study of institution. *Ration. Soc.* **2002**, *14*, 7–53.
36. Modica, S.; Tallon, J.; Rustichini, A. Unawareness and bankruptcy: A general equilibrium model. *Econ.Theory* **1998**, *12*, 259–292.
37. Laffont, J.; Martimort, D. *The Theory of Incentives: The Principal-Agent Model*; Princeton University Press: Princeton, NJ, USA, 2002.
38. Halpern, J.Y.; Rêgo, L.C. Extensive games with possibly unaware players. Mathematical Social Sciences, forthcoming, 2012.
39. Rêgo, L.C.; Halpern, J.Y. Generalized solution concepts in games with possibly unaware players. *Int. J. Game Theory* **2012**, *41*, 131–155.
40. Camerer, C.; Ho, T.; Chong, J. A cognitive hierarchy theory of one-shot games. *Q. J. Econ.* **2004**, *119*, 861–898.
41. Blume, L.; Brandenburger, A.; Dekel, E. Lexicographic probabilities and equilibrium refinements. *Econometrica* **1991**, *59*, 81–98.
42. Radner, R. Bounded rationality, indeterminacy, and the theory of the firm. *Econ. J.* **1996**, *106*, 1360–1373.
43. Debreu, G. A social equilibrium existence theorem. *Proc. Natl. Acad. Sci. USA* **1952**, *38*, 886–893.

44. Glicksberg, I. A further generalization of the Kakutani fixed point theorem, with application to Nash equilibrium points. *Proc. Natl. Acad. Sci. USA* **1952**, *3*, 170–174.
45. Fan, K. Fixed-point and minimax theorems in locally convex topological linear spaces. *Proc. Natl. Acad. Sci. USA* **1952**, *38*, 121–126.