Universidad de San Andrés

Departamento de Economía

Licenciatura en Economía

Asimetría de Información e Inteligencia Artificial: Una aplicación de modelos mixtos

Autores: María Guastavino y Joaquín Lardizabal

Legajos: 30106 y 30277

Mentores de Tesis: Christian Ruzzier y Lucía Quesada

Buenos Aires, 27 de julio de 2022

Abstract

El siguiente trabajo busca aplicar los conceptos de modelos de asimetría de información a un caso particular, a partir del cual se pueden derivar conclusiones y aprendizajes. En este trabajo se va a plantear la resolución de un modelo mixto de Adverse Selection seguido por Moral Hazard, donde además incluimos un costo catástrofico futuro que depende del accionar del agente hoy. Este adicional busca entender cómo se ven afectados los incentivos del agente hoy cuando su accionar puede tener repercusiones (y en este caso, catastróficas) a futuro. El principal hallazgo del trabajo es respecto al tipo de contratos que se encuentran en condiciones de asimetría de información (también denominado Second Best): en caso de tener a ambos agentes activos, las transferencias que reciben son exactamente las mismas. En otras palabras, con los agentes activos el contrato que se encuentra es un contrato *pooling*. Por otro lado, también encontramos que cuando no necesariamente se quiere tener a todos los agentes activos, esto solo se puede realizar si se dejan afuera los agentes de costos más altos, definición que le vamos a otorgar a los tipos de agentes que sufren más la catástrofe.

1. Introducción

La Inteligencia Artificial es un tema muy contemporáneo, algo que parecía una fantasía hace unas décadas y que hoy ya es una realidad. Muchos nos imaginamos un futuro conviviendo con máquinas que nos ayuden en nuestras tareas del día a día, y que, en esencia, estén hechas para hacer de nuestra vida una más simple.

En este trabajo de tesis lo que nos proponemos es realizar modelos incorporando la teoría aprendida en Economía de la Información, para poder simular las interacciones entre las personas a la hora de construir una Inteligencia Artificial. Vamos a considerar varios factores, como las múltiples dimensiones de los jugadores, o el uso de potenciales catástrofes futuras en el accionar de hoy, para construir nuestros modelos y poder derivar conclusiones respecto a esta interacción.

Fue leyendo el libro *Superintelligence: Paths, Dangers, Strategies* (2017) de Nick Bostrom que surgió la inspiración para hacer este trabajo. En el mismo se habla de los peligros de una Superinteligencia Artificial (SI), como podemos ver en las películas *The Terminator* (1984) con Skynet o *I, Robot* (2004) con Sonny y los demás robots, y de los cuidados que se deberían tener al construir una máquina con habilidades superiores a las de los humanos en todo aspecto. Al crear este tipo de tecnologías, siempre existe la posibilidad de que la SI, al tener tanto alcance y 'poder' (en el sentido de acceso ilimitado a todo tipo de tecnología e información, una capacidad de procesamiento infinitamente superior a la humana), pueda convertirse en un riesgo para la humanidad. Este riesgo podría ser, por ejemplo, el control de los humanos o, directamente, la completa exterminación de la especie.

Pero antes de hablar directamente del modelo, queremos hacer un repaso sobre cómo podría llegarse a una situación como la que tenemos en mente. Los orígenes de la inteligencia artificial se remontan al siglo pasado, donde el término fue oficialmente introducido por Marvin Minsky y John McCarthy (2006) en el *Dartmouth Summer Research Project on Artificial Intelligence*. La métrica que se usaba en ese entonces para discernir entre máquina inteligente o no es el conocido *Test de Turing*: si un humano está interactuando con otro humano y con una máquina, y no puede distinguir a la máquina del humano, entonces a esa máquina se la considera inteligente (Haenlein y Kaplan, 2019) Obviamente han habido grandes avances desde esa época, y hoy día la cuestión no es cuándo sino cómo las máquinas van a pasar a una posición de líderes y toma de decisiones.

Esto está sucediendo a tal escala, que el *California Management Review* (2019) tiene una edición especial sobre IA. Haenlein y Kaplan (2019) hacen un gran trabajo resumiendo los puntos principales de esta edición, donde se enfocan en cómo la IA afectará las relaciones entre empresas y empleados, el mercado laboral en general, el mecanismo de toma de decisiones interno de las empresas, y la relación entre firmas y consumidores. En todas estas se proponen esquemas de colaboración entre humanos e IAs, con algunos puntos donde las IA deben tomar las decisiones por sí solas, sin interacción con humanos. La transición a una vida *AI-first*, entonces, está cada vez más cerca.

Uno de los puntos que nos cuestionamos con este trabajo es el por qué se le dio tanto poder a la IA, como es que el ser humano confía tanto en ellas. Durante nuestra investigación para este trabajo, nos encontramos con el paper de Schniter et al. (2019), que plantea un experimento para probar las diferencias en las interacciones entre humanos y robots. Realizan un experimento en el que humanos interactúan con otros humanos o robots en una serie de juegos, y encuentran que el trato de humanos a humanos y robots es similar, a pesar de saber que estan lidiando con un ser no emocional como es un *software.* Creemos que esto es aún más relevante para nuestro trabajo, porque no solo están las inteligencias artificiales entrando a nuestra vida, sino que lo hacen con nuestro consentimiento, aprobación y confianza. Por otro lado, Huang et al. (2021) y Young et al. (2021) también proponen distintos mecanismos mediante los que la IA puede funcionar en el sector público. Huang et al. (2021) muestra que los gerentes son más propensos a confiar y utilizar IA que los empleados en el sector público, mientras que Young et al. (2021) propone mecanismos mediante los que la IA puede prevenir daño innecesario en la administración pública.

Hemos visto que estamos transicionando a un mundo donde las IA pasan a tener posiciones de poder. Ahora bien, esto conlleva ciertos riesgos. Acemoglu (2021) está especialmente preocupado

con este tópico, como muestra en su trabajo en proceso *"Harms of AI"*, donde especifica una serie de problemas que prevé pueden suceder si las inteligencias artificiales siguen siendo aplicadas sin control ni regulación, como lo está siendo en el curso que se está tomando actualmente. Hay una salvedad que se hace en el trabajo que consideramos importante, y es que "Los costos que he modelado no son inherentes a la naturaleza de estas tecnologías de IA, sino dependen de como esta nueva plataforma tecnológica se está usando y desarrollando para empoderar corporaciones y gobiernos contra ciudadanos y trabajadores." (Acemoglu, 2021: pág 46 - nuestra traducción) Las IA, entonces, no son malas por sí solas, sino que depende de la forma en la que estén construidas y de la aplicación que se les de.

¿Qué podemos hacer cuando las IA se vuelvan tan certeras que no podamos distinguirlas de un ser humano? ¿Qué límites tienen las corporaciones a usarlas en nuestra contra? No hace falta proyectar respuestas a estas preguntas, existe un caso muy reciente donde un ingeniero de Google afirma que un proyecto de IA que estan desarrollando se ha vuelto verdadermanete inteligente. La empresa no estuvo de acuerdo, por lo que el empleado fue despedido y su reclamo fue desestimado, incluso sugiriendole que busque ayuda psiquiátrica (Grant y Metz, 2022)

Viendo entonces el rol que pueden llegar a tomar las inteligencias artificiales, la capacidad de toma de decisiones y posiciones de poder que pueden alcanzar en un futuro cercano, como nosotros confiamos en ellas y les otorgamos este poder, y que no está habiendo regulaciones sobre su aplicación, el escenario donde una IA tome el control por nuestra mano y desate un caos apocalíptico, entonces, no es una idea tan descabellada.

Entiendase bien: como Acemoglu, nosotros también consideramos que los peligros de las IA no son inherentes a ellas, sino que pueden ser productos de una falla en el desarrollo o implementación de estas. Es por esto que una solución que planteamos es el siguiente modelo teórico, donde el problema radica en la construcción y desarrollo de una IA en un modelo de principal-agente con información asimétrica seguido de riesgo moral.

El trabajo sigue con el desarrollo del modelo teórico, explicando los componentes y el timing del mismo, seguidos por la resolución de los modelos de First Best y Second Best. Luego interpretamos los modelos, y por último llegamos a conclusiones y proponemos futuros pasos.

Adelantando los resultados obtenidos en el trabajo, vamos a ver que para este caso en particular, sucede que hay dos tipos de contratos de Second Best posibles: un contrato donde ambos agentes se encuentran activos, y otro donde solo un tipo de agente se encuentra activo.

En el primer contrato, donde ambos agentes reciben transferencias, vamos a encontrar que esta solución es de un contrato *pooling*, donde las transferencias se diferencian según el resultado del trabajo (es decir, una Inteligencia Artificial buena o mala técnicamente) a modo de incentivar el esfuerzo alto, pero no se encuentran diferencias según el tipo del agente.

Por el otro lado, para el segundo contrato donde solo un tipo de agente se encuentra activo (en otras palabras, contrato con Shutdown), encontramos que este caso solo es factible cuando dejamos afuera al que más le pesa la catástrofe, o, en otras palabras, el agente que más costos tiene. Esto se debe a que en caso de querer mantener activo a este agente, y entonces tener que ofrecer una transferencia que cubra sus costos, no hay forma de evitar que el otro agente (que no le importa tanto la catástrofe, por lo que tiene costos menores) no quiera mentir sobre su tipo y entrar al contrato.

2. El modelo

El modelo teórico que tenemos en mente para el trabajo describe el conflicto de interés que hay a la hora de construir y programar una Superinteligencia Artificial. El agente va a ser el programador al cual se le designó la creación de la Superinteligencia Artificial. Este también podría ser un grupo de programadores, ingenieros y científicos a los cuales se les delegó la tarea, pero para simplificar

el ejercicio vamos a considerar que todo este grupo puede ser representado en un único agente. Por otro lado, el principal va a estar representado por una empresa contratadora. El principal, al igual que el agente, también podría ser toda la humanidad, ya que esta tiene en sus intereses que el agente construya la Superinteligencia Artificial de la mejor manera posible (y que no busque destruir a la humanidad como forma de salvarla, por ejemplo), pero vamos a considerar que los intereses de la empresa contratadora están alineados a los de la humanidad. Además, considerar al principal como una empresa, en vez de toda la humanidad, facilita el planteo de las funciones de utilidad de ambos actores.

A la hora de armar la SI, vamos a considerar que el agente tiene dos dimensiones: por un lado, vamos a tener que considerar el tipo del agente, y por el otro, el agente tiene una decisión sobre su nivel de esfuerzo (que, para simplificar, la decisión va a ser si se esfuerza o no). El tipo del agente, en este modelo en particular, va a hacer referencia a la actitud frente a su trabajo y la responsabilidad que tiene sobre el mismo, ya que vamos a clasificar a los tipos de agente en aquellos que les importa la posible catástrofe ($\overline{\theta}$) y los que no les importa ($\underline{\theta}$). Vamos a considerar que la clasificación de un agente en cualquiera de los tipos tiene una probabilidad conocida:

$$\begin{cases} P(\theta = \overline{\theta} = 1) = b \\ P(\theta = \underline{\theta} = 0) = 1 - b \end{cases}$$

Para facilitar los análisis a lo largo del trabajo, una consideración que vamos a tener es que $\overline{\theta} = 1$ y que $\underline{\theta} = 0$. Aunque muchas veces vamos a resolver los modelos usando las θ, otras veces usaremos los valores 0 y 1. Esta simplificación nos va a resultar particularmete útil para algunos desarrollos, para interpretar los resultados de los modelos y definir condiciones respecto a ciertos parámetros.

Respecto al nivel de esfuerzo (e), la decisión que toma el agente afecta el resultado de su trabajo, pero también se debe tener en cuenta que hay elementos a la hora de la construcción de la máquina que el agente no puede controlar y que vamos a considerar como aleatorios. Por eso, la SI que programe el agente (el resultado de su trabajo, que matemáticamente vamos a denominar SI) puede ser buena (SI_h) o mala (SI_l) a nivel tecnológico, y este resultado depende en parte del esfuerzo que realice el agente, y en parte de los shocks externos aleatorios (por ejemplo, puede haber un mal resultado técnico aunque el agente decida esforzarse debido a shocks externos). Eso sí, vamos a considerar que con un esfuerzo alto ($e = 1$) las probabilidades de tener un mejor resultado son mayores que si el agente decide no esforzarse ($e = 0$). Una observación para este modelo, es que este esfuerzo va a ser independiente del tipo del agente. Entonces, como vamos a considerar que con un esfuerzo alto las probabilidades de tener un mejor resultado son mayores que si el agente decide no esforzarse, tendremos:

$$\begin{cases} P(SI = SI_h | e = 1) = \pi_1 \\ P(SI = SI_h | e = 0) = \pi_0 \end{cases}$$

Notemos que la probabilidad de que la SI salga bien técnicamente depende positivamente del esfuerzo, por lo que $\pi_1 > \pi_0$. También vamos a utilizar más adelante la expresión para el diferencial entre estas probabilidades: $\Delta\pi = \pi_1 - \pi_0 > 0$. Además, resulta útil notar que $1 - \pi_1$ no es lo mismo que π_0. Lo que sí es, es la probabilidad de que la SI salga mal a pesar de esforzarse.

Eso sí, el esfuerzo al agente le genera un costo personal de $\psi(e)$ independiente del tipo de agente, y para cada valor que puede tomar e, los costos se expresan de la siguiente manera:

$$\begin{cases} \psi(e = 1) = \psi \\ \psi(e = 0) = 0 \end{cases}$$

Pero el esfuerzo del agente afecta a más de una dimensión de la máquina. En otras palabras, el esfuerzo que realiza el agente tiene un efecto en la estructura y uso tecnológico de la SI, pero también tiene un impacto en la 'moralidad' de la máquina (es decir, si decide o no destruir a la humanidad). A la hora de elegir el nivel de esfuerzo, el agente puede dedicar más tiempo y esfuerzo para hacer que esta tenga la menor cantidad de errores técnicos y que el riesgo de que la SI se torne en contra de la raza humana sea menor (lo cual le cuesta al agente) o puede no hacerlo y directamente llevarse el pago que le ofrece la empresa por su trabajo. En otras palabras, el esfuerzo que realiza el agente va a afectar tanto al resultado del trabajo (SI buena o mala técnicamente) y a la probabilidad del apocalipsis (SI buena o mala moralmente).

Ese costo apocalíptico (al cual vamos a denominar a, un valor independiente en este modelo, que toma un valor necesariamente positivo ($a \geq 0$)) va a representar la posibilidad de una catástrofe si es que la SI decide exterminar a toda la humanidad. Este costo va a afectar a todos los jugadores (tanto agente como el principal), porque efectivamente si se extermina a toda la humanidad, nadie va a estar a salvo. Las probabilidades de pagar este costo se van a ver afectadas por el esfuerzo también, ya que si un agente se esfuerza más a la hora de armar la SI, seguramente pueda crear más exitosamente un programa para darle 'moralidad' a la máquina (o simplemente un programa para que no destruya a la humanidad o algo similar). Este factor es independiente del tipo, y la notación que vamos a usar es la siguiente:

$$\begin{cases} P(a|e=1) = v_1 \\ P(a|e=0) = v_0 \end{cases}$$

En este caso, al esforzarse el agente está reduciendo la probabilidad de enfrentarse al costo a, por lo que necesariamente $v_1 < v_0$. En la notación vamos a expresar este diferencial como: $\Delta v = v_0 - v_1 > 0$.

Otro aspecto importante a tener en cuenta es que vamos a suponer que el principal no puede condicionar el contrato que le ofrece al agente al hecho de que suceda o no la catástrofe y la Inteligencia Artificial decida destruir a toda la humanidad, dado que en este caso toda la humanidad se verá exterminada.

Adicionalmente, y respecto al efecto del costo catastrófico, vamos a considerar que este afecta de manera distinta a los tipos de agente dependiendo de su tipo. Como mencionamos anteriormente, vamos a tener dos tipos: uno que le importa evitar lo máximo posible el evento catastrófico porque la catástrofe le pesa mucho ($\overline{\theta}$) y otro que le importa menos el evento catastrófico ($\underline{\theta}$), posiblemente porque considera que para cuando suceda la catástrofe, él no va a seguir estando. Por ende, vamos a considerar que en nuestro modelo la θ de cada tipo se encuentra junto a a (la catástrofe, en nuestra ecuación) para representar el distinto peso que tiene a para cada tipo de agente. Relacionado a este efecto, como al primer tipo de agente le afecta más la catástrofe que al segundo tipo, debe suceder que $\overline{\theta}a > \underline{\theta}a$, por ende $\overline{\theta} > \underline{\theta}$ (y, en efecto, $\Delta\theta = \overline{\theta} - \underline{\theta} > 0$).

También vamos a considerar que ambos actores son neutrales al riesgo, aunque el agente tiene un límite de lo que puede pagar ($t \geq l \geq 0$).

Por último, y para seguir refinando nuestro análisis y poder derivar más conclusiones, más adelante vamos a calcular el pago esperado del principal.

Un detalle de las transferencias, que vamos a tener en cuenta a la hora de armar los modelos, es que estas van a ser distintas respecto al tipo del agente ($\underline{\theta}$ o $\overline{\theta}$) y al resultado obtenido del trabajo del agente (siendo t_h para un resultado bueno y t_l para uno malo) gracias al principio de revelación. En efecto, vemos que las transferencias entonces pueden tomar alguna de las siguientes formas:

$$\begin{cases} t(SI = SI_h|\theta = \overline{\theta} = 1) = \overline{t_h} \\ t(SI = SI_h|\theta = \underline{\theta} = 0) = \underline{t_h} \\ t(SI = SI_l|\theta = \overline{\theta} = 1) = \overline{t_l} \\ t(SI = SI_l|\theta = \underline{\theta} = 0) = \underline{t_l} \end{cases}$$

Otro detalle sobre las transferencias planteadas, es que en algunos casos las vamos a agrupar por resultado o por el tipo de agente. En este caso, para simplificar la notación, vamos a escribir una t genérica que aplique para ese caso. Por ejemplo, si agrupamos a las t según los tipos de agente, vamos a tener $\overline{t}$ y $\underline{t}$, que equivaldrían a $\overline{t} = \overline{t_h} = \overline{t_l}$ y a $\underline{t} = \underline{t_h} = \underline{t_l}$. Por el otro lado, si agrupamos según el resultado, tendríamos t_h y t_l, donde $t_h = \underline{t_h} = \overline{t_h}$ y $t_l = \underline{t_l} = \overline{t_l}$.

Finalmente, para este modelo vamos a considerar que los hechos transcurren de la siguiente manera:

1. El agente ve su tipo (actitud frente a la catástrofe), pero el principal no
2. El principal le ofrece un contrato al agente
3. El agente acepta o rechaza el contrato
4. El agente elige el nivel de esfuerzo que va a realizar
5. Se realiza el trabajo
6. Se observa el resultado del trabajo por todos los jugadores (resultado se ve afectado por el esfuerzo del agente y por la influencia de la naturaleza)
7. Se ejecuta el contrato (se realizan las transferencias)

2.1. First Best

Antes de resolver el problema con asimetría de información, queremos ver cuál es el resultado al cual se llega en el caso que no haya asimetría, en otras palabras: el resultado eficiente.

En el caso de First Best, las únicas restricciones que vamos a considerar son las de participación (IR) y las de limited liability (LL), ya que como el principal puede ver el esfuerzo y el tipo del agente, no es necesaria la condición de compatibilidad de incentivos (el agente se va a esforzar y tampoco va a mentir sobre su tipo). En este caso, vamos a tener dos restricciones de participación, una por cada tipo de agente.

Una de las consideraciones que vamos a tener para el modelo es que el agente no tiene riqueza previa o activos al entrar a la relación como para hacer un pago al principal, por lo que $l = 0$. Otro supuesto que vamos a realizar es que es eficiente para el principal que el agente realice esfuerzo alto (supuesto cuyas condiciones para que se cumpla vamos a verificar más adelante). Por ende, el modelo resultante es el siguiente:

$$Max_{t_h;t_l}\ \pi_1(s(SI_h) - t_h) + (1 - \pi_1)(s(SI_l) - t_l) - v_1 a$$

Sujeto a:

$$IR : \pi_1 t_h + (1 - \pi_1)t_l - \psi - v_1\theta a \geq 0$$

$$LL_h : t_h \geq 0$$

$$LL_l : t_l \geq 0$$

Antes de empezar a resolver el modelo y conseguir el contrato óptimo de First Best, hay varias consideraciones que debemos resolver. La primera es verificar que el principal quiera inducir esfuerzo alto. Para eso, el beneficio social esperado de inducir esfuerzo debe ser mayor al de no inducirlo. El beneficio del principal dada una realización del contrato y el costo apocalíptico es $s(SI) - t - a$, y el del agente es $t - \psi(e) - a\theta$. El beneficio social es, entonces, $w = s(SI) - a(1 + \theta) - \psi(e)$, que es la suma de los beneficios planteados. Vemos dos variables en la formula, por lo que tenemos w que dependen de θ y e, resultando la siguiente, que es el beneficio social esperado:

$$w(\theta; e) = \pi(e)SI_h + (1 - \pi(e))SI_l - v(e)a(1 + \theta) - \psi(e)$$

Con $e = 0$; tenemos $\pi_0 SI_h + (1-\pi_0)SI_l - v_0 a(1+\theta)$, que simplificamos a $SI_l + \pi_0 \Delta SI - v_0 a(1+\theta)$

Con $e = 1$; tenemos $\pi_1 SI_h + (1-\pi_0)SI_l - v_1 a(1+\theta)$, que simplificamos a $SI_l + \pi_1 \Delta SI - v_1 a(1+\theta) - \psi$

Para encontrar que $e = 1$ es eficiente, deseable, necesitamos que $w(\theta; 0) \leq w(\theta; 1)$, independientemente del θ que encontremos ($\forall\theta$). Resolviendo esta inecuación, llegamos a la siguiente condición:

$$\psi \leq \Delta v a(1 + \theta) + \Delta SI \Delta\pi$$

De esta condición vemos que esforzarse es eficiente cuando el costo del esfuerzo (ψ) es menor al beneficio del esfuerzo, que viene por dos fuentes: un aumento en la probabilidad de tener una SI técnicamente buena ($\Delta\pi\Delta SI$); y por una reducción en la probabilidad de sfurir un costo apocalíptico, que afecta tanto al principal como al agente ($\Delta v a(1 + \theta)$). También es fácil notar que esta condición depende de a. Cuanto más grande es el costo apocalíptico, mayor es el rango de valores de ψ para los cuales el esfuerzo es eficiente.

Como $\overline{\theta} > \underline{\theta}$; si la condición anterior se cumple para $\underline{\theta}$, se cumple para $\overline{\theta}$ también. Entonces, para que el principal quiera inducir esfuerzo alto, necesitamos que: $\psi \leq \Delta v a(1 + \underline{\theta}) + \Delta SI \Delta\pi$

En caso de considerar los θ como $\overline{\theta} = 1$ y $\underline{\theta} = 0$, obtenemos la siguiente condición:

$$\psi \leq \Delta v a + \Delta SI \Delta\pi$$

En segundo lugar, debemos verificar que sea eficiente para la sociedad construir una Super Inteligencia Artificial. Para eso, y sabiendo que esfuerzo alto es deseable bajo la anterior condición, el beneficio social esperado debería ser positivo. Es decir, $w(\theta; 1) \geq 0$. Resolviendo de la misma manera que para la condición anterior, otra vez teniendo en cuenta que $\overline{\theta} \geq \underline{\theta}$; encontramos que la IA es deseable si:

$$a \leq \frac{SI_l + \pi_1 \Delta SI - \psi}{v_1(1 + \overline{\theta})}$$

Esta condición se deriva de que el costo social de la SI tiene que ser menor al beneficio social esperado. Lo que nos termina mostrando, entonces, es que el costo apocalíptico esperado tiene que ser menor al beneficio social de la SI menos el costo del esfuerzo de los agentes, lo cuál le pone un techo al impacto de la catástrofe en los jugadores (tanto el principal como el agente), porque

si el costo apocalíptico es muy grande (por ejemplo, tiende el infinito), el costo social va a ser demasiado grande y no hay beneficio que lo compense.

En caso de considerar θ como $\overline{\theta} = 1$ y $\underline{\theta} = 0$, obtenemos la siguiente condición sobre a:

$$a \leq \frac{SI_l + \pi_1 \Delta SI - \psi}{2v_1}$$

Esta condición le pone un límite superior al costo de la catástrofe en nuestro problema, por lo que debemos tenerlo en cuenta para la resolución. En otras palabras, cuando el costo apocaliptico (a) es muy grande, lo mejor que pueden hacer todos para ahorrárselo es no construir la Super Inteligencia Artificial. Vamos a suponer que la condición anterior sobre a se cumple en lo que sigue del trabajo.

Un último cálculo que podemos realizar antes de encontrar el contrato First Best, es el costo de implementación del principal en este caso. Esto lo podemos hacer sin encontrar las transferencias porque, en First Best, el principal va a cubrir únicamente los costos del agente, que en este caso se conforman por el costo del esfuerzo y el costo apocalíptico. En otras palabras, el costo esperado del principal en First Best es:

$$C^{FB} = \psi + (b\overline{\theta} + (1-b)\underline{\theta})av_1$$

Teniendo $\overline{\theta} = 1$ y $\underline{\theta} = 0$, nuestro costo queda reducido a:

$$C^{FB} = \psi + bav_1$$

Esto se debe a que solo el agente de tipo $\overline{\theta}$ sufre el costo apocalíptico (costo que el principal debe cumplir), por lo que el principal siempre debe cubrir el costo del esfuerzo (ψ) y tiene una probabilidad igual a b de cubrir el costo apocalíptico ($v_1 a$), probabilidad que se traduce a la probabilidad de tener un agente con costo apocalíptico.

Ahora, para encontrar la solución de First Best, comenzamos activando la IR:

$$\pi t_h + (1-\pi_1)t_l - \psi - v_1 a\theta = 0$$

$$\pi_1 t_h + (1-\pi_1)t_l = \psi + v_1 a\theta$$

De esta manera podemos encontrar un conjunto de transferencias que soluciona el problema de First Best: cualquier par de t que satisfaga esta condición y las de Limited Liability (es decir, que las transferencias sean mayores o iguales a cero). Hay más de un menú de transferencias para First Best dado que nos encontramos en una situación donde tanto el agente como el principal son neutrales al riesgo, entonces hay una gran variedad de asignaciones de riesgo que podrían ser factibles.

Un ejemplo de conjunto de t que satisface estas restricciones es:

$$\begin{cases} \overline{t^*} = \overline{t_h} = \overline{t_l} = \psi + v_1 \overline{\theta} a \\ \underline{t^*} = \underline{t_h} = \underline{t_l} = \psi + v_1 \underline{\theta} a \end{cases}$$

Este conjunto de transferencias, el principal se hace cargo de todo el riesgo relacionado a que la realización de la Inteligencia Artificial (SI) sea técnicamente buena. Además, en este caso también sucede que el principal, al conocer el tipo del agente, es capaz de diferenciarlos y ofrecer el contrato que le corresponde a cada uno, siendo así capaz de extraerle toda la renta del agente de la manera más barata posible. Otra observación sobre este resultado es que $\overline{t^*} > \underline{t^*}$.

Otros dos conjuntos de transferencias que encontramos resuelven el problema son tomando el valor de algunas de las transferencias como igual a cero, y despejando el valor de la otra transferencia. En este caso, se diferenciaría según el resultado y el tipo del agente:

$$\begin{cases} t_h^* = \overline{t_h} = \underline{t_h} = \dfrac{\psi + v_1 \theta a}{\pi_1} \\ t_l^* = \overline{t_l} = \underline{t_l} = 0 \end{cases}$$

$$\begin{cases} t_h^* = \overline{t_h} = \underline{t_h} = 0 \\ t_l^* = \overline{t_l} = \underline{t_l} = \dfrac{\psi + v_1 \theta a}{1 - \pi_1} \end{cases}$$

Como mencionamos en el planteo del modelo, una consideración que vamos a usar para facilitar los análisis es usar $\overline{\theta} = 1$ y $\underline{\theta} = 0$. En el caso de First Best, teniendo en cuenta esta consideración sobre los θ, nos quedarían los siguientes conjuntos de solución:

$$\begin{cases} \overline{t^*} = \overline{t_h} = \overline{t_l} = \psi + v_1 a \\ \underline{t^*} = \underline{t_h} = \underline{t_l} = \psi \end{cases}$$

$$\begin{cases} \overline{t_h^*} = \dfrac{\psi + v_1 a}{\pi_1} \\ \underline{t_h^*} = \dfrac{\psi}{\pi_1} \\ t_l^* = \overline{t_l} = \underline{t_l} = 0 \end{cases}$$

$$\begin{cases} \overline{t_l^*} = \dfrac{\psi + v_1 a}{1 - \pi_1} \\ \underline{t_l^*} = \dfrac{\psi}{1 - \pi_1} \\ t_h^* = \overline{t_h} = \underline{t_h} = 0 \end{cases}$$

En todos los casos vemos que el principal lo que hace es compensar los costos del agente, a veces con algunas ponderaciones, pero estos costos son distintos según el tipo de agente. Para los de menor costo ($\underline{\theta}$) solo se compensa el valor del esfuerzo, porque a ellos no les pesa la catástrofe (dado que $\underline{\theta} = 0$). Por el otro lado, para los de mayor costo ($\overline{\theta}$) también se les debe compensar por el costo de la catástrofe. Más adelante haremos un análisis más exhaustivo de este resultado.

2.2. Second Best

Considerando todos los datos brindados, que el esfuerzo ni el tipo de agente son observables ni verificables por el principal y que el agente no tiene renta como para pagarle al principal, el problema que nos queda para resolver es el siguiente:

$$Max_{\overline{t_h};\overline{t_l};\underline{t_h};\underline{t_l}}\ b[\pi_1(s(SI_h)-\overline{t_h})+(1-\pi_1)(s(SI_l)-\overline{t_l})]+(1-b)[\pi_1(s(SI_h)-\underline{t_h})+(1-\pi_1)(s(SI_l)-\underline{t_l})]-v_1 a$$

Sujeto a:

$$\overline{IC_{as}} : \pi_1\overline{t_h} + (1-\pi_1)\overline{t_l} - \psi - v_1\overline{\theta}a \geq Max_{e\in 0;1}\ [\pi(e)\underline{t_h} + (1-\pi(e))\underline{t_l} - \psi(e) - v(e)\overline{\theta}a]$$

$$\underline{IC_{as}} : \pi_1\underline{t_h} + (1-\pi_1)\underline{t_l} - \psi - v_1\underline{\theta}a \geq Max_{e\in 0;1}\ [\pi(e)\overline{t_h} + (1-\pi(e))\overline{t_l} - \psi(e) - v(e)\underline{\theta}a]$$

$$\overline{IC_{mh}} : \pi_1\overline{t_h} + (1-\pi_1)\overline{t_l} - \psi - v_1\overline{\theta}a \geq \pi_0\overline{t_h} + (1-\pi_0)\overline{t_l} - v_0\overline{\theta}a$$

$$\underline{IC_{mh}} : \pi_1\underline{t_h} + (1-\pi_1)\underline{t_l} - \psi - v_1\underline{\theta}a \geq \pi_0\underline{t_h} + (1-\pi_0)\underline{t_l} - v_0\underline{\theta}a$$

$$\overline{IR} : \pi_1\overline{t_h} + (1-\pi_1)\overline{t_l} - \psi - v_1\overline{\theta}a \geq 0$$

$$\underline{IR} : \pi_1\underline{t_h} + (1-\pi_1)\underline{t_l} - \psi - v_1\underline{\theta}a \geq 0$$

$$\overline{LL_h} : \overline{t_h} \geq 0$$

$$\underline{LL_h} : \underline{t_h} \geq 0$$

$$\overline{LL_l} : \overline{t_l} \geq 0$$

$$\underline{LL_l} : \underline{t_l} \geq 0$$

Como podemos ver, hay cuatro restricciones de compatibilidad de incentivos. Esto se debe a que el agente en este problema tiene dos dimensiones, una respecto a su tipo y otra respecto al esfuerzo. Por ende, el principal debe garantizar para cada tipo que este se esfuerce (IC_{mh}) y que, además, no mientan respecto a su tipo (IC_{as}).

Con este modelo en mente, lo primero que vamos a hacer es usar nuestra solución original de First Best para ver qué sucede con las restricciones de Second Best. Este es un ejercicio que nos va a ayudar a identificar las restricciones que tenemos que activar para encontrar la solución de Second Best. Como en el First Best encontramos más de un conjunto de transferencias que son solución, vamos a utilizar las siguientes transferencias para hacer el ejercicio:

$$\begin{cases} \overline{t^*} = \overline{t_h} = \overline{t_l} = \psi + v_1\overline{\theta}a \\ \underline{t^*} = \underline{t_h} = \underline{t_l} = \psi + v_1\underline{\theta}a \end{cases}$$

Para simplificar la resolución de las restricciones y la interpretación de las condiciones, vamos a utilizar nuestra consideración $\overline{\theta} = 1; \underline{\theta} = 0$ de ahora en adelante.

- $\overline{IC_{as}}$ con $e = 1$: a partir de $\pi_1(\psi + v_1 a) + (1-\pi_1)(\psi + v_1 a) - \psi - v_1 a \geq \pi_1\psi + (1-\pi_1)\psi - \psi - v_1 a$ obtenemos $0 \geq -\Delta\theta v_1 a$, que se cumple siempre.

- $\overline{IC_{as}}$ con $e = 0$: a partir de $\pi_1(\psi+v_1 a)+(1-\pi_1)(\psi+v_1 a)-\psi-v_1 a \geq \pi_0\psi+(1-\pi_0)\psi-v_0 a$ obtenemos una condición (factible) de a: $a \geq \frac{\psi}{v_0}$

- $\underline{IC_{as}}$ con $e = 1$: a partir de $\pi_1\psi + (1-\pi_1)\psi - \psi \geq \pi_1(\psi + v_1 a) + (1-\pi_1)(\psi + v_1 a) - \psi$ llegamos a un absurdo: $0 \geq v_1 a$

- $\underline{IC_{as}}$ con $e = 0$: a partir de $\pi_1\psi + (1-\pi_1)\psi - \psi \geq \pi_0(\psi + v_1 a) + (1-\pi_0)(\psi + v_1 a)$ llegamos a una condición sobre a: $a \geq \frac{\psi}{\Delta v}$

- $\overline{IC_{mh}}$: a partir de $\pi_1(\psi+v_1 a)+(1-\pi_1)(\psi+v_1 a)-\psi-v_1 a \geq \pi_0(\psi+v_1 a)+(1-\pi_0)(\psi+v_1 a)$ llegamos a la condición sobre a: $a \geq \frac{\psi}{\Delta v}$

- $\underline{IC_{mh}}$: a partir de $\pi_1\psi+(1-\pi_1)\psi-\psi \geq \pi_0\psi+(1-\pi_0)\psi$ llegamos a un absurdo: $\psi \leq 0$

A partir de los absurdos encontrados en la restricción de $\underline{IC_{as}}$ y $\underline{IC_{mh}}$, concluimos que el agente siempre va a querer mentir sobre su tipo si debe esforzarse, y además, siempre va a preferir no esforzarse en caso de solo poder desvíarse en esfuerzo (no en tipo). Por ende, estas son las resticciones que vamos a tener que activar para encontrar el conjunto solución de Second Best.

Ahora sí, para encontrar la resolución del Second Best, vamos a activar $\underline{IC_{as}}$, $\underline{IC_{mh}}$, $\overline{LL_l}$ y $\underline{LL_l}$. Estas últimas las vamos a activar porque consideramos que todas las transferencias sobre buenos resultados (t_h) no van a ser mayores que las transferencias por malos resultados (t_l). Esta es una consideración que vamos a validar más adelante.

A partir de las restricciones primero notamos rápidamente que:

$$\begin{cases} \underline{LL_l} : \underline{t_l} = 0 \\ \overline{LL_l} : \overline{t_l} = 0 \end{cases}$$

Utilizamos estos valores y los reemplazamos en las otras dos restricciones que vamos a activar. Es así como, a partir de la IC_{mh} encontramos $\underline{t_h} = \frac{\psi}{\Delta\pi}$; y a partir de la IC_{as} encontramos el valor de $\overline{t_h} = \frac{\psi}{\Delta\pi}$

Destacamos entonces el conjunto de transferencias que son solución para este modelo:

$$\begin{cases} \underline{t_l} = 0 \\ \overline{t_l} = 0 \\ \underline{t_h} = \dfrac{\psi}{\Delta\pi} \\ \overline{t_h} = \dfrac{\psi}{\Delta\pi} \end{cases}$$

Y con este resultado en mente, chequeamos que cumplan las restricciones que no activamos:

- $\overline{IC_{as}}$ con esfuerzo: $\frac{\pi_1}{\Delta\pi}(\psi) - \psi - v_1 a \geq \frac{\pi_1}{\Delta\pi}(\psi) - \psi - v_1 a$ que notamos siempre se cumple.

- $\overline{IC_{as}}$ sin esfuerzo: $\frac{\pi_1}{\Delta\pi}(\psi) - \psi - v_1 a \geq \frac{\pi_0}{\Delta\pi}(\psi) - v_0 a$ y se reduce a $\Delta v a \geq 0$, que también se cumple siempre.

- $\overline{IC_{mh}}$: $\frac{\pi_1}{\Delta\pi}(\psi) - \psi - v_1 a \geq \frac{\pi_0}{\Delta\pi}(\psi) - v_0 a$ de la que encontramos $\Delta v a \geq 0$, que se cumple siempre con nuestros supuestos.

- $\overline{IR}$: $\frac{\pi_1}{\Delta\pi}(\psi) - \psi - v_1 a \geq 0$, de la que encontramos la condición $a \leq \pi_1 \frac{\psi}{(v_1 \Delta\pi)}$

- $\underline{IR}$: $\frac{\pi_1}{\Delta\pi}(\psi) - \psi \geq 0$, que se cumple siempre bajo nuestros supuestos.

- $\overline{LL_h} = \underline{LL_h}$: $\frac{(\psi)}{\Delta\pi} \geq 0$, que se cumple siempre bajo nuestros supuestos.

Un elemento que queremos verificar es cuál es la condición límite sobre a. Como podemos ver, a lo largo de la resolución del modelo, encontramos más de una condición sobre el costo apocalítico, y debemos chequear cuál es la más demandante para poder establecer los límites y también analizar qué sucede en caso de violarse.

Empezamos notando que las cuatro restricciones que fueron activadas no tienen condiciones de a, como tampoco las tienen las que se cumplen siempre ($\overline{IC_{as}}$ con y sin esfuerzo e $\overline{IC_{mh}}$.) Tenemos entonces la condición extraida de $\overline{IR}$:

$$a \leq \pi_0 \frac{\psi}{v_1 \Delta\pi}$$

Con esta condición en mente, nuestro problema de Second Best queda resuelto con el menú de transferencias presentado anteriormente: $(\underline{t_l} = \overline{t_l} = 0; \underline{t_h} = \overline{t_h} = \frac{\psi}{\Delta\pi})$

Por último, y como paso adicional, queremos mostrar cuáles son las condiciones sobre $\overline{t_h}$ tal que se cumplan las restricciones de IC_{as}, adelantando el resultado, como forma de mostrar que en este caso particular es imposible tener un contrato separador (es decir, que se hagan distinciones sobre los tipos de los agentes), que no sea el contrato de Shutdown. En otras palabras, un contrato separador con ambos agentes activos.

Empezamos el análisis suponiendo $\underline{LL_l}, \overline{LL_l}$ activas, por lo que $\underline{t_l} = \overline{t_l} = 0$

A partir de la restricción de compatibilidad de incentivos de Moral Hazard para el agente de menor costo ($\underline{IC_{mh}}$) obtenemos la siguiente condición: $\underline{t_h} = \frac{\psi}{\Delta\pi}$.

Luego, miramos las restricciones de Adverse Selection para el mismo agente ($\underline{IC_{as}}$), con y sin esfuerzo, y obtenemos resultados interesantes:

- para $e = 1$: encontramos $\pi_1 \underline{t_h} - \psi \geq \pi_1 \overline{t_h} - \psi$, por ende concluímos que: $\overline{t_h} \leq \underline{t_h}$

- para $e = 0$: encontramos $\pi_1 \underline{t_h} - \psi \geq \pi_0 \overline{t_h}$, y reordenando llegamos a que: $\overline{t_h} \leq \frac{\pi_1}{\pi_0} \underline{t_h} - \frac{\psi}{\pi_0}$

Si seguimos jugando con esas condiciones encontradas sobre t_h, y consideramos el resultado encontrado arriba de $\underline{t_h} = \frac{\psi}{\Delta\pi}$, obtenemos finalmente:

$$\overline{t_h} \leq \frac{\psi}{\Delta\pi}$$

Ahora, si hacemos el mismo procedimiento pero con la otra condición de compatibilidad de incentivos de Adverse Selection, $\overline{IC_{as}}$ con y sin esfuerzo, obtenemos:

- para $e = 1$ encontramos $\pi_1 \overline{t_h} - \psi - v_1 a \geq \pi_1 \underline{t_h} - \psi - v_1 a$, por ende concluímos que: $\overline{t_h} \geq \underline{t_h}$

- para $e = 0$: encontramos $\pi_1 \overline{t_h} - \psi - v_1 a \geq \pi_0 \underline{t_h} - v_0 a$, y reordenando llegamos a que: $\overline{t_h} \geq \frac{\pi_0}{\pi_1} \underline{t_h} + \frac{\psi}{\pi_1} - \frac{\Delta v_a}{\pi_1}$

Ya observando las condiciones que encontramos de $\underline{IC_{as}}$ y $\overline{IC_{as}}$ para $e = 1$, nos damos cuenta que necesariamente debe pasar que $\underline{t_h} = \overline{t_h}$. Entonces, si se quiere tener a ambos agentes activos, necesariamente debe suceder que:

$$\overline{t_h} = \underline{t_h} = \frac{\psi}{\Delta\pi}$$

Con este análisis, mostramos que no puede haber un contrato separador (es decir, un contrato que ofrezca transferencias diferentes a los agentes) sin que sea el caso de Shutdown, donde alguno de los agentes se encuentra inactivo. Gracias a esto, y validando lo que habíamos mencionado tras la primer resolución de Second Best, el contrato que encontramos es *pooling*. Este contrato, entonces, es el único bajo la condición encontrada para a.

Ahora, si no buscamos un contrato donde ambos agentes se encuentran activos y, además, establecemos otros valores de parámetros a este problema, podemos encontrar una solución que satisfaga el problema con estos valores. Para eso, hacemos uso de un contrato con Shutdown. En este, el principal le ofrece un contrato que alguno de los tipos no va a aceptar para descartarlo. En nuestro caso, elegimos intentar descartar al tipo con costo catastrófico alto ($\overline{\theta}$ que recibe $\overline{t_h}$ y $\overline{t_l}$), por lo que nos queda un agente que puede decidir entre dos niveles de esfuerzo distintos, por lo tanto con una sola dimensión de variabilidad ($\underline{t_h}$ y $\underline{t_l}$). Para simplificar la notación sin perder generalidad, vamos a renombrar a las transferencias al agente como t_h y t_l.

Con estos tipos de agente en mente, el problema se reduce a un solo tipo y Moral Hazard. Tenemos tres condiciones a nuestra dispocisión: LL_l, IC e IR. Activamos LL_l, obteniendo:

$$t_l = 0$$

Continuamos activando $\underline{IC_{mh}}$ y resolvemos utilizando $\underline{\theta} = 0$:

$$\pi_1 t_h - \psi - v_1\underline{\theta}a = \pi_0 t_h - v_0\underline{\theta}a$$
$$t_h = \frac{\psi}{\Delta\pi}$$

Verificamos en IR:

$$\pi_1(\frac{\psi}{\Delta\pi}) - \psi - v_1\underline{\theta}a \geq 0$$
$$0 \geq -\pi_0\psi$$

Además, debemos verificar que el tipo $\overline{\theta}$ no quiera mentir con su tipo. Para eso, debe suceder que la utilidad de mentir sea menor a la de decir la verdad. Tenemos que demostrar esto tanto para las ecuaciones con esfuerzo como sin esfuerzo:

$$e = 1 : \pi_1\underline{t_h} + (1 - \pi_1)\underline{t_l} - \psi - v_1 a$$
$$e = 0 : \pi_0\underline{t_h} + (1 - \pi_0)\underline{t_l} - v_0 a$$

Para que ambas sean negativas, debe darse que:

$$e = 1 : a > \frac{\pi_0\psi}{v_1\Delta\pi}$$

$$e = 0 : a > \frac{\pi_0 \psi}{v_0 \Delta\pi}$$

Como deben cumplirse ambas ecuaciones, se simplifica el análisis si una de las condiciones es suficiente para la otra. Y, efectivamente, $a > \frac{\pi_0 \psi}{v_1 \Delta\pi}$ es condición suficiente para $a > \frac{\pi_0 \psi}{v_0 \Delta\pi}$. De esta manera entendemos que esta condición debe satisfacerse para que el contrato de Shutdown sea un equilibrio.

Al cumplirse siempre la restricción para IR, y con la condición sobre a en mente, podemos afirmar que este menú de transferencias es solución al problema de Shutdown, y también al problema de Second Best:

$$\begin{cases} \overline{t_h} = \overline{t_l} = \underline{t_l} = 0 \\ \underline{t_h} = \dfrac{\psi}{\Delta\pi} \end{cases}$$

Consideramos (y demostraremos) que este es el único contrato de Shutdown posible, y que no se puede hacer para el agente de costos bajos ($\underline{\theta}$). Intuitivamente, y observando los resultados obtenidos a lo largo del paper, notamos que las transferencias que el principal realiza a los agentes lo que hacen es, principalmente, cubrir los costos de los mismos. Estas transferencias, debido a las restricciones de información, se pueden ver levemente alteradas, pero en rangos generales siguen esta regla.

Esa es la principal razón por la cual creemos que hacer un contrato de Shutdown para dejar afuera al agente de costos altos ($\underline{\theta}$), mientras que se tiene activo al de costos alto ($\overline{\theta}$) no es posible. Al agente de costos bajos siempre le va a resultar atractiva la transferencia del otro agente, porque esta debería ser más alta para cubrir los costos mayores de $\overline{\theta}$, que no solo incluyen los costos del esfuerzo, sino que también incluye el costo apocalíptico (recordemos que consideramos $\underline{\theta} = 0$, por lo que el costo a se ve ponderado por un valor que equivale a cero. Igualmente, el argumento se mantiene si se usan θs genéricos, porque $\underline{\theta} \leq \overline{\theta}$).

Igualmente, esto también lo podemos ver tomando $\underline{t_h} = \underline{t_l} = 0$ y viendo si se satisfacen o se violan las restricciones.

Vamos a seguir el mismo procedimiento que con el Shutdown para el agente de costos altos, donde el problema se reduce a un único tipo (en este caso $\overline{\theta}$) y Moral Hazard y tenemos tres condiciones a nuestra dispocisión: LL_l, IC e IR. al igual que en el ejercicio anterior, activamos LL_l y obtenemos:

$$t_l = 0$$

Luego, activando $\overline{IC_{mh}}$ y resolvemos utilizando $\overline{\theta} = 1$:

$$\pi_1 t_h - \psi - v_1 \overline{\theta} a = \pi_0 t_h - v_0 \overline{\theta} a$$
$$t_h = \frac{\psi - \Delta v a}{\Delta\pi}$$

Al verificar que se cumpla la IR, obtenemos una condición sobre a. En otras palabras, si a no satisface la siguiente ecuación, el agente con costos altos no va a querer entrar en el contrato, haciéndolo nulo:

$$a \leq \frac{\psi(\frac{\pi_1}{\Delta\pi} - 1)}{\Delta v \frac{\pi_1}{\Delta\pi} + v_1}$$

Esta condición puede cumplirse en ciartas condiciones, pero vamos a chequear que, dado que el caso que se cumpla y el agente de costos altos quiere entrar al contrato, entonces no suceda que un agente de costos bajos también quiera entrar haciéndose pasar por el otro. En otras palabras, vamos a chequear la restricción de compatibilidad de incentivos de Adverse Selection para el agente de costos bajos ($\underline{IC_{as}}$).

Planteando y resolviendo esa inecuación con el contrato encontrado, llegamos que el agente de costos bajos no va a mentir sobre su tipo (y hacerse pasar por el de costos altos), siempre y cuando se cumpla la siguiente condición sobre a:

$$a \geq \frac{\psi}{\Delta v}$$

Entonces, y por último, lo que vamos a chequear es que estas dos condiciones de a se puedan satisfaceer en simultáneo. Si la respuesta es no, entonces siempre que el agente de costos altos quiera participar del contrato (porque haya casos donde no participa y todas las transferencias pasan a ser nulas), el de costos bajos se va a ver motivado para entrar y mentir sobre su tipo, haciéndose pasar por el otro. Si la respuesta es si, entonces efectivamente encontramos un contrato factible de Shutdown donde el agente de costos bajos se queda afuera, sin ningun motivo para mentir, y el de costos altos participa.

Entonces, lo que sigue es verificar que ambas condiciones de a se puedan dar en simultáneo:

$$\frac{\psi(\frac{\pi_1}{\Delta\pi} - 1)}{\Delta v \frac{\pi_1}{\Delta\pi} + v_1} \geq a \geq \frac{\psi}{\Delta v}$$

Que, al sacar el a, dejando únicamente las condiciones límite, y resolviendo la inecuación, llegamos al siguiente absurdo:

$$v_0 \leq 0$$

Entonces, a modo de resumen y retomando lo que habíamos hipotetizado previamente, este tipo de contrato de Shutdown donde el agente de costos bajos queda afuera es imposible en este modelo, porque siempre va a tener incentivos a mentir sobre su tipo y hacerse pasar por el costos altos.

3. Interpretación de Resultados

Como mencionamos a lo largo de la resolución de los modelos planteados, nos encontramos en un caso donde, dependiendo de la configuración de los parámetros, encontramos distintos conjuntos de transferencias que son solución para nuestro modelo. En esta sección vamos a analizar e interpretar los resultados obtenidos.

En primer lugar, para el caso de First Best, encontramos un conjunto de transferencias que son solución del problema. Uno de ellos es el siguiente:

$$\begin{cases} \overline{t^*} = \overline{t_h} = \overline{t_l} = \psi + v_1 a \\ \underline{t^*} = \underline{t_h} = \underline{t_l} = \psi \end{cases}$$

En este caso, las transferencias son una para cada tipo y cubren los valores esperados de esfuerzo para ambos. Esto se debe a que el principal puede ver el tipo y el nivel de esfuerzo de todos los tipos de agente (lo cual los obliga a realizar un esfuerzo alto y decir la verdad respecto a su tipo). Adicionalmente, observamos que la transferencia es independiente del resultado (que también depende de factores externos aleatorios), porque en este caso es el principal el que se hace responsable de todo el riesgo y cubre los costos esperados de cada tipo (costo del esfuerzo y costo apocalíptico para cada tipo.)

Pero en First Best también podemos encontrar soluciones donde el principal da una transferencia igual a cero para cierto resultado, sin importar el tipo, y otra transferencia positiva para el resultado opuesto:

$$\begin{cases} \overline{t_h^*} = \dfrac{\psi + v_1 a}{\pi_1} \\ \underline{t_h^*} = \dfrac{\psi}{\pi_1} \\ t_l^* = \overline{t_l} = \underline{t_l} = 0 \end{cases}$$

$$\begin{cases} \overline{t_l^*} = \dfrac{\psi + v_1 a}{1 - \pi_1} \\ \underline{t_l^*} = \dfrac{\psi}{1 - \pi_1} \\ t_h^* = \overline{t_h} = \underline{t_h} = 0 \end{cases}$$

En estos casos, el principal al tener toda la información sobre los niveles de esfuerzo y los tipos de agente, puede nuevamente diferenciar entre ellos y darles distintas transferencias a cada tipo según el resultado. Es importante notar que para las transferencias que son distintas de cero, el valor para cada tipo es distinto, ya que depende de θ. Entonces, aunque parezca que son las mismas transferencias para ambos tipos independiente del resultado, hay una pequeña variación. Además, destacamos que en cualquier solución la transferencia esperada de $\overline{\theta} > \underline{\theta}$ porque tiene que ser compensado por un costo mayor gracias al costo apocalíptico.

Ahora, observemos el resultado obtenido en Second Best, considerando el caso que ambos tipos de agente se encuentran activos:

$$\begin{cases} \underline{t_l} = 0 \\ \overline{t_l} = 0 \\ \underline{t_h} = \dfrac{\psi}{\Delta\pi} \\ \overline{t_h} = \dfrac{\psi}{\Delta\pi} \end{cases}$$

Con estas transferencias lo que podemos ver es que, dada la asimetría de información del accionar y del tipo del agente, el principal se ve forzado a realizar transferencias positivas únicamente cuando el resultado técnico de la Super Inteligencia Artificial es bueno. Vemos muy claro que el principal ofrece un contrato donde se hace una diferencia de las transferencias según el resultado (SI_h o SI_l) y no el tipo del agente. Lo que podemos intuir a partir de este resultado es que si el

principal se ve forzado a satisfacer Limited Liability y, al mismo tiempo, incentivar esfuerzo alto, no puede ofrecer contratos que dependan del tipo del agente. Este tipo de contratos se denominan *contratos pooling*, donde el principal cede flexibilidad sobre el tipo de contratos que ofrece (las transferencias no dependen de todas las dimensiones, como resultados y tipo de agente, si no que solo dependen de una de ellas), a cambio de tener un conjunto de transferencias que satisfacen todas las restricciones. Al solo tener las transferencias como herramientas para satisfacer todas las restricciones, se ve forzado a hacer *pooling* con tal de satisfacerlas todas.

En caso de no tener a ambos tipos de agente activos, lo que encontramos fue un contrato para Second Best donde se realiza un Shutdown. En otras palabras, uno de los tipos de agente se encuentra inactivo. Para nuestro caso particular, el agente que vamos a tomar como inactivo es el que tiene los costos más altos, que en este caso es $\overline{\theta}$. Esto se debe a que este agente, al importarle tanto la catástrofe, tiene costos que el otro tipo de agente no posee. Es más, al considerar $\overline{\theta} = 1$ y $\underline{\theta} = 0$, esto lo vemos con claridad (el costo de la catástrofe es $av(e)\theta$).

El contrato, en este caso de Shutdown, es el siguiente:

$$\begin{cases} \overline{t_h} = \overline{t_l} = \underline{t_l} = 0 \\ \underline{t_h} = \dfrac{\psi}{\Delta\pi} \end{cases}$$

Para que suceda el Shutdown, explicamos que $\overline{\theta}$ no debe tener incentivos a mentir, y que para eso suceda encontramos la condición $a > \frac{\pi_0 \psi}{v_1 \Delta\pi}$.

Esta condición sobre a es complementaria a la que encontramos antes. Tenemos Shutdown si a es lo suficientemente alto como para que el costo del tipo alto sea más caro mantenerlo que no pagarle.

Volviendo al contrato con Shutdown, vemos que en la solución, el agente de costo alto queda inactivo, es decir, excluido del contrato, mientras que el de costo bajo solo recibe un transferencia positiva en caso de buenos resultados.

También demostramos que este contrato de Shutdown es el único factible, dado que nunca se va a poder realizar un contrato donde el agente de costos bajos quede inactivo y el de costos altos activo. Esto se debe principalmente porque las transferencias del conjunto solución suelen seguir un patrón de cubrir los costos de los agentes. Por ende, las transferencias para los agentes de costos altos son muy atractivas para los de costos bajos en un contexto donde no reciben nada ($\underline{t_h} = \underline{t_l} = 0$).

Otro aspecto que es clave y muy claro es que hay una diferencia respecto a las transferencias de First Best. Dadas las restricciones de información esto era de esperarse. El principal, considerando que no sabe ni el tipo del agente ni el esfuerzo que se realiza, se ve forzado a realizar un contrato con un costo de implementación mas alto para garantizar el mejor resultado posible dadas las nuevas restricciones.

Para poder verificar esto, hicimos un rápido análisis de costos esperados que enfrenta el principal en cada uno de los casos (First y Second Best), para ver este desvío causado por información asimétrica. Entonces, vamos a caluclar los costos esperados para el principal usando la esperanza de t (cuál es la transferencia esperada que el principal esperaría pagar considerando las distintas probabilidades de tipo y resultado):

$$E(t) = b\pi_1\overline{t_h} + b(1-\pi_1)\overline{t_l} + (1-b)\pi_1\underline{t_h} + (1-b)(1-\pi_1)\underline{t_l}$$

Como calculamos anteriormente, el costo esperado de First Best es el siguiente:

$$C^{FB} = \psi + (b\overline{\theta} + (1-b)\underline{\theta})av_1$$

Costo que, con $\overline{\theta} = 1$ y $\underline{\theta} = 0$, queda reducido a:

$$C^{FB} = \psi + bav_1$$

Habiendo obtenido el resultado de los costos esperados para First Best, ahora nos queda obtener los de Second Best para así poder analizar las distorsiones:

$$\begin{cases} \underline{t_l} = 0 \\ \overline{t_l} = 0 \\ \underline{t_h} = \dfrac{\psi}{\Delta\pi} \\ \overline{t_h} = \dfrac{\psi}{\Delta\pi} \end{cases}$$

$$E(t) = b\pi_1 \frac{\psi}{\Delta\pi} + (1-b)\pi_1 \frac{\psi}{\Delta\pi}$$

Desarrollando, llegamos al costo esperado del principal de Second Best:

$$C^{SB} = \pi_1 \frac{\psi}{\Delta\pi}$$

Teniendo ambas esperanzas, podemos compararlas restando la de Second Best a la de First Best. De esta comparación, verificamos que el costo esperado de First Best sea menor al de Second Best:

$$\frac{\pi_1\psi}{\Delta\pi} \geq \psi + v_1 ab$$

De la que resulta:

$$a \leq \frac{\pi_0}{v_1 b \Delta\pi}$$

Entonces, siempre que se cumpla la condición sobre a que encontramos, se va a cumplir que los costos esperoados del principal en First Best son menores a los de Second Best. Y, en nuestro caso, como $b < 1$, y habíamos concluido anteriormente a partir de la restricción de participación del agente de costo bajo (en el ejercicio de Second Best), que $a \leq \frac{\pi_0\psi}{\Delta\pi v_1}$, entonces necesariamente se va a cumplir la inecuación planteada. En otras palabras, el costo de First Best siempre va a ser menor al de Second Best, por lo que encontramos que con asimetría de informacion se llega a contratos subóptimos. Estos contratos subóptimos se deben a que, con asimetría de información, hay casos donde, dados los valores de los parámetros, inducir esfuerzo alto es eficiente, pero el

principal prefiere no inducirlo.

Por último, considerando que en Second Best según el valor de a, el principal va a tener que ofrecer distintos contratos. Cuando el valor de a es más pequeño, el contrato que se va a ofrecer es uno donde ambos tipos de agentes activos, mientras que a crece, el contrato que se ofrece es uno de Shutdown. A continuación, queremos identificar cuál de ellos es mejor en costos esperados para el principal.

Por lo tanto, vamos a calcular el costo esperado del principal de Shutdown, considerando que tenemos el siguiente contrato:

$$\begin{cases} \overline{t_h} = \overline{t_l} = \underline{t_l} = 0 \\ \underline{t_h} = \dfrac{\psi}{\Delta\pi} \end{cases}$$

$$E(t) = (1-b)\pi_1 \frac{\psi}{\Delta\pi}$$

Por ende, los costos esperados del principal en Second Best con contrato de Shutdown son:

$$C^{SB}_{shutdown} = (1-b)\pi_1 \frac{\psi}{\Delta\pi}$$

Rápidamente notamos que, en relación a los costos del principal donde ambos agentes están activos, los costos en este caso son menores. Esto tiene sentido si consideramos que las transferencias distinas a cero que tiene que hacer el principal son menores: en Shutdown solo hay transferencias cuando el agente de costos altos hace un buen trabajo, mientras que en el contrato con ambos agentes activos, las transferencias se hacen siempre que se haga un buen trabajo, independiente del tipo de agente.

4. Conclusión

Como pudimos observar a lo largo del trabajo realizado, tener asimetría de información resulta en distorsiones en las interacciones entre los distintos tipos de agente, y cuanta más asimetría hay (en este caso tenemos asimetría de información en una mayor cantidad de dimensiones), más se van a distorsionar los resultados obtenidos.

Los modelos planteados en este trabajo no resuelven, ni buscan resolver, preguntas pertinentes sobre la construcción de una Super Inteligencia Arificial, como, por ejemplo: ¿Cuál es la forma óptima de construir una SI? O ¿Qué hay que hacer para asegurarse que la construcción de la SI se haga de la mejor manera posible para no terminar en una situación de Skynet? Sino que, en los modelos planteados, se busca establecer cuánto uno se desvía de lo óptimo, en nuestro caso el First Best, cuando no se puede acceder a toda la información a la hora de armar los contratos.

Es más, hasta pudimos observar que a pesar de tener un incentivo futuro (y un incentivo muy fuerte como lo puede ser la destrucción de toda la humanidad), el principal es incapaz de llegar al contrato óptimo de First Best cuando no tiene la información completa del tipo y del accionar del agente. Es verdad que algunas restricciones de los modelos se relajan a partir de este costo adicional, pero en ninguno de los casos se pudo llegar a la solución óptima que se da únicamente

cuando hay información perfecta. Recordemos que encontramos una condición sobre a que nos pone un techo al valor de esta catástrofe. De representar un costo muy alto para la humanidad, es preferible para la sociedad no desarrollarla.

Además, pudimos observar que no es posible llegar a una respuesta única, sino que los resultados en el modelo varían según la configuración de los parámetros. De esta manera, podemos observar distintos conjuntos de transferencias que satisfacen las restricciones del modelo dependiendo de la configuración en la que estemos. Vimos, por ejemplo, como funcionaría un caso con Shutdown, y justificamos por qué es racional dejar de lado al tipo que más le importa la catástrofe: porque conlleva un costo mas alto.

El caso planteado en este trabajo, donde consideramos un programador que debe esforzarse para que la Super Inteligencia Artificial que está construyendo funcione técnica y moralmente, es simplemente una forma de ilustrar el problema de Moral Hazard y Adverse Selection combinados, cuando hay costos futuros que dependen del accionar del agente hoy.

Como detallamos al comienzo del trabajo, este modelo podría complejizarse más aún, dejando de lado supuestos que simplificaron la resolución del modelo. Por ejemplo, se podría considerar al esfuerzo como continuo en vez de binario, se podrían considerar tipos de agente agente y/o principales aversos al riesgo, o también el costo catastrófico como una función no lineal. Este trabajo sirve como una primera aproximación a todas las complejidades que se le pueden agregar a los típicos modelos mixtos de Adverse Selection y Moral Hazard, pero todavía hay espacio para expandir los horizontes y ver los efectos de las distintas alteraciones.

Esperamos que, a la hora de crear, construir y programar una Super Inteligencia Artificial, los encargados de la misma se encuentren alineados en qué significa realizar un buen trabajo. A fin de cuentas, como pudimos ver aquí, alinear tanto las capacidades técnicas de la máquina con su moralidad y la empatía con la raza humana, es de interés colectivo.

Bibliografía

Acemoglu, D. (2021). “Harms of AI.” SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3922521.

Bostrom, N. (2017). Superintelligence : Paths, Dangers, Strategies. Oxford: Oxford University Press, Cop.

Grant, Nico, y Metz, Cade. (2022). "Google Sidelines Engineer Who Claims Its A.I. Is Sentient."New York Times, June 12. https://www.nytimes.com/2022/06/12/technology/google-chatbot-ai-blake-lemoine.html (accessed July 11, 2022)

Haenlein, M., and Kaplan, A.. (2019). “A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence.” California Management Review 61 (4): 5–14. https://doi.org/10.1177/0008125619864925.

Huang, H., Kyoung-Cheol (Casey) K., Young, M., and Bullock, J. (2021). “A Matter of Perspective: Differential Evaluations of Artificial Intelligence between Managers and Staff in an Experimental Simulation.” Asia Pacific Journal of Public Administration 44 (1): 47–65. https://doi.org/10.1080/23276665.2021.1945468.

Laffont, J and Martimort, D. (2003). The Principal Agent Model : The Economic Theory of Incentives. Cheltenham, Uk ; Northhampton, Ma, Usa: E. Elgar Pub.

Laffont, J. (1995). “Regulation, Moral Hazard and Insurance of Environmental Risks.” Journal of Public Economics 58 (3): 319–36. https://doi.org/10.1016/0047-2727(94)01488-a.

McCarthy, J., Minsky, M. L., Rochester, N., Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. AI Magazine, 27(4), 12. https://doi.org/10.1609/aimag.v27i4.1904.

Schniter, E., T. W. Shields, and D. Sznycer. (2020). “Trust in Humans and Robots: Economically Similar but Emotionally Different.” Journal of Economic Psychology 78 (June): 102253. https://doi.org/10.1016/j.joep.2020.102253.

Young, M., Himmelreich, J., Bullock, J., Kyoung-Cheol, K. (2021). “Artificial Intelligence and Administrative Evil, Perspectives on Public Management and Governance". Oxford Academic, Volume 4, Issue 3, September 2021, Pages 244–258. https://doi.org/10.1093/ppmgov/gvab006.