UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE CIENCIAS BIOLÓGICAS
DEPARTAMENTO DE MATEMÁTICA APLICADA (BIOMATEMÁTICA)

APRENDIZAJE Y ESTABILIZACIÓN DE COMPORTAMIENTOS ALTRUISTAS EN SOCIEDADES DE AGENTES AUTÓNOMOS

TESIS DOCTORAL

Javier Zamora Romero

MADRID, 1997

UNIVERSIDAD COMPLUTENSE DE MADRID
DEPARTAMENTO DE MATEMÁTICA APLICADA (BIOMATEMÁTICA)

APRENDIZAJE Y ESTABILIZACIÓN DE COMPORTAMIENTOS ALTRUISTAS EN SOCIEDADES DE AGENTES AUTÓNOMOS

Javier Zamora Romero

Memoria presentada
para optar al grado de Doctor

Vº Bº Directores de la tesis

Fdo. José del Rocío Millán Ruiz

Fdo. Antonio Murciano Cespedosa

MADRID, 1997

Agradecimientos

Mi primer agradecimiento va destinado a los directores de esta tesis. José del Rocío Millán Ruíz ha aportado su experiencia científica, su inmensa capacidad de trabajo y unas grandes dosis de paciencia para, a pesar de mis demoras y de la distancia kilométrica que nos separa, hacer posible la llegada a buen fin de este proyecto. De Antonio Murciano Cespedosa he recibido su contagioso entusiasmo y su ayuda incondicional. Además de director, ha desempeñado los papeles de gran amigo y gran compañero. Mi más sincero agradecimiento para ambos.

Desde mi incorporación al Departamento de Matemática Aplicada, he tenido la fortuna de poder contar con numerosos apoyos de mis compañeros. Quiero agradecer a la directora del mismo, Mª. Cristina Martínez Calvo, su confianza y su generosa gestión encaminada siempre a posibilitar la realización de éste y otros trabajos. Mi gratitud también hacia Alberto Pérez de Vargas por sus constantes muestras de inquitud, no sólo científica hacia el trabajo desarrollado, sino también personal hacia todos los que formamos parte del entorno del departamento. Quisiera también mostrar mi agradecimiento hacia Julio Alonso Fernández. Mi contacto inicial con el mundo de la investigación se lo debo a él. Sus sinceros ánimos y su acertada crítica han sido, y serán siempre, motivo de agradecimiento. Quiero dar también las gracias a todos los compañeros y amigos que, de una forma u otra, han contribuido al desarrollo de esta tesis.

La última línea de estos agradecimientos la quiero dedicar a Maripi, mi mujer, que pese a sufrir la tesis como sólo la mujer de un doctorando sabe, ha permanecido prestando el aliento que sólo un doctorando sabe apreciar y agradecer. Gracias.

A Maripi,
a mis padres.

INDICE

Capítulo 1

Introducción

1.1 Motivación de la tesis y marco conceptual.

La Inteligencia Artificial (IA) es un área de investigación que pretende sintetizar procedimientos inteligentes para implementarlos en máquinas que realicen trabajos que requieren de procesamiento inteligente de la información. La visión más tradicional de la IA fundamenta sus herramientas en la adquisición y representación del conocimiento, y se denomina comúnmente IA simbólica. Esta escuela tienen un carácter marcadamente secuencial que consta esencialmente de dos pasos: las entidades del problema real se representan internamente mediante símbolos y manipulación de estos símbolos, a través de mecanismos de inferencia, para la búsqueda de soluciones eficientes a problemas complejos. Este abordaje es adecuado cuando el dominio del problema está perfectamente definido y estático. Pero en ocasiones, los problemas a los que se debe enfrentar la IA son aquellos que se denominan "situados". En ellos, la máquina está interaccionando directamente con el dominio del problema, recabando la información pertinente mediante sus sensores y actuando sobre el ambiente con sus sistemas motores. Esto convierte el dominio del problema en algo dinámico, modificado continuamente por las acciones de la máquina, que debe tomar sus decisiones en un tiempo limitado a partir de una información frecuentemente incompleta y con ruido. Ante este tipo de situaciones, el abordaje tradicional de la IA simbólica, con un funcionamiento secuencial, muestra soluciones poco

eficientes. Se hace imprescindible entonces la búsqueda de nuevas metodologías de trabajo que ofrezcan soluciones ante esta nueva clase de problemas.

Una inagotable vía de inspiración es la observación del comportamiento animal. Los seres vivos son un claro ejemplo de agentes *situados* que en su actividad diaria se enfrentan a problemas complejos que requieren toma de decisiones tales como alimentarse o huir, recolectar o descansar, etc. mostrando frente a ellos soluciones eficientes. Sus respuestas se realizan en tiempo real, con información parcial del problema y frecuentemente ruidosa pero resultan adecuadas para la supervivencia del individuo. El paralelismo que existe tanto en la problemática como en los objetivos de ambos contextos, artificial y natural, hace indicado el estudio de las soluciones adoptadas por los seres vivos para su aplicación a la robótica. De este camino de ida y vuelta pueden obtenerse dos tipos de beneficios. Por un lado, mediante la obtención de modelos fieles que recojan las principales variables explicativas de dichos fenómenos, es posible realizar simulaciones y experimentos controlados que ayuden a conocer la evolución de los sistemas biológicos en escalas de tiempo sólo abordables bajo las coordenadas que la instrumentación informática actual posibilita. Por otro lado, tras la validación de estos modelos de comportamiento, éstos pueden implementarse en los sistemas artificiales con el fin de obtener soluciones similares a las biológicas en problemas que hayan demostrado ser también similares. Este flujo de información entre la naturaleza y la inteligencia artificial ha dado ya interesantes frutos. Los modelos de aprendizaje en redes neuronales artificiales, los modelos de optimización mediante algoritmos genéticos y la programación genética entre otros, son ejemplos de esta vía de trabajo.

La presente tesis se encuadra en este marco de observación y modelización de la naturaleza en busca de propuestas para solucionar problemas en la robótica inteligente. En particular se pretende resolver el grave problema de estabilidad que la robótica colectiva afronta cuando un colectivo de agentes puede seleccionar entre varias alternativas de comportamiento cooperativo siendo alguna de ellas el comportamiento altruista. Desarrollaremos en el siguiente apartado el problema propuesto con más detalle.

1.2 Planteamiento del problema.

Uno de los principales objetivos de la robótica inteligente es el denominado de *optimalidad*. Con él se pretenden alcanzar rendimientos óptimos o cuasi-óptimos en la ejecución de tareas. En el caso de la robótica colectiva, para la consecución de dicho objetivo se propone la inclusión en los sistemas multiagente de estrategias de comportamiento social. Estas estrategias sociales pueden ser, entre otras, la cooperación,

competición, especialización y altruismo. Definimos un comportamiento social como *altruista* cuando suponga un beneficio para el receptor del acto y un coste para el actor del mismo. En el mismo sentido, se define un comportamiento social como *egoísta* cuando el actor de dicho comportamiento obtenga beneficios a costa de pérdidas sufridas por otro agente.

En determinadas circunstancias ambientales, la presencia de comportamientos altruistas incrementa la eficiencia del colectivo, aunque por definición, provoca un detrimento de la eficacia individual del altruista. Ahora bien, la eficacia de estos comportamientos depende del ambiente de trabajo y éste puede variar o ser desconocido a priori. Es por tanto imposible preprogramar, en un sistema multiagente, un repertorio de comportamientos cooperativos válido para cualquier ambiente. Por esta razón se hace necesaria la inclusión en el sistema de mecanismos de adaptación. Con ellos, los agentes modificarán su comportamiento seleccionando aquel que brinde mejores rendimientos en el ambiente en que esté situado. Si la evaluación de las distintas alternativas de comportamiento se realiza con información local a los agentes, esto es, medidas de la eficacia individual, y los agentes tienden a maximizar esta medida, elegirán la estrategia más rentable, desplazando otras estrategias, individualmente menos rentables, del acervo de comportamientos de la población. En el caso de los comportamientos sociales en los que se puede elegir entre cooperar o defraudar, la mayor rentabilidad a corto plazo corresponde a los comportamientos egoístas (defraudadores). Por esta razón, bajo las condiciones expuestas hasta el momento, nada impediría que un colectivo de agentes con comportamiento cooperativo altruista fuese invadido por el comportamiento egoísta. El límite será la extinción del carácter altruista y su sustitución por el comportamiento egoísta. Esto conduce al conocido problema de estabilidad del altruismo.

Además de esta inestabilidad, se tiene un efecto paradójico en relación con el rendimiento del sistema. El rendimiento que un agente obtiene al mostrar una de las dos alternativas del comportamiento cooperativo -cooperar o no hacerlo-, depende del comportamiento que muestre el resto de los agentes. En concreto, el rendimiento de la estrategia egoísta depende de la presencia o no de altruistas en la población. Es decir, si existen individuos altruistas en la población, un individuo con comportamiento egoísta puede aprovechar los esfuerzos cooperativos del resto sin pagar ningún coste por ello. Hemos visto que la inestabilidad del altruismo conducirá a su desaparición del sistema, invadido por la estrategia no cooperadora. En este punto, el rendimiento de los agentes egoístas será incluso menor que el obtenido si hubiesen asumido los costes de la cooperación. El efecto conjunto de la inestabilidad y de este comportamiento paradójico del rendimiento llevará al sistema a adoptar soluciones no óptimas.

La presente tesis propone un nuevo principio, inspirado en el comportamiento animal, que permite a cada agente individual aprender comportamientos altruistas estables a fin de obtener rendimientos óptimos del colectivo.

1.3 Enunciado de la tesis y principales aportaciones.

La presente memoria enuncia la siguiente tesis "*el altruismo recíproco es una estrategia aprendible capaz de estabilizar comportamientos altruistas en sociedades de agentes autónomos con capacidad de aprendizaje individual, aún sin conocer el rendimiento colectivo* ".

Esta tesis afronta, mediante la estrategia extraída de la naturaleza denominada reciprocidad, el problema existente en la robótica colectiva de estabilidad de la cooperación costosa para el individuo. El uso conjunto de la reciprocidad y de sistemas de adaptación a partir de información local de los robots, demuestra ser efectivo en el mantenimiento del altruismo en sistemas multiagente que trabajan en entornos situados.

En esta memoria se presenta un sistema multiagente desarrollando una tarea de recogida de objetos en ambientes situados. Este sistema cuenta con las siguientes propiedades:

1. Autonomía: los agentes toman sus decisiones de forma autónoma a partir de la información percibida por sus sensores y de su estado interno.
2. Control no centralizado: el control del equipo de trabajo no depende de ningún controlador central, sino que emerge de las interacciones entre los agentes.
3. Robustez: el sistema resiste ante perturbaciones provocadas por malfuncionamiento de alguno de los agentes, adición de ruido en las señales y en las lecturas sensoriales, etc.
4. Cooperación: los agentes realizan su trabajo de forma cooperativa incrementando el rendimiento colectivo. La cooperación está fundamentada en comunicación explícita, no dirigida y de bajo coste.
5. Altruismo: los agentes cooperan con los demás, aún a costa de pérdidas de eficacia individual.

El objetivo fundamental de la tesis es hacer compatible, en un sistema multiagente con las características propuestas, los siguientes aspectos del dilema optimalidad-estabilidad.

1. Optimalidad: el sistema deberá alcanzar rendimientos óptimos (o cercanos al óptimo) en el ambiente en que se desenvuelve.
2. Adaptabilidad: los agentes deberán modificar sus comportamientos eligiendo los más adecuados para las distintas situaciones ambientales. Cada agente hará esta elección de forma independiente utilizando información exclusivamente local.

3. Flexibilidad: ante cambios ambientales, el sistema deberá ser capaz de revertir su configuración de comportamientos para adaptarse a la nueva situación.
4. Estabilidad: pese al carácter local de la información procesada por los agentes, la cooperación deberá ser estable frente a la invasión de estrategias de comportamiento no cooperativas.

La consecución de estos objetivos provocará que el colectivo de agentes estabilice la estrategia de comportamiento cooperativo que obtenga rendimientos óptimos (o cuasi-óptimos) para cualquier ambiente.

Las principales aportaciones de esta tesis son:

1. Desarrollo de la arquitectura AREA.

   Se desarrolla una arquitectura de agente autónomo, basada en comportamientos, que incluye los mecanismos necesarios para la implementación de la estrategia de reciprocidad. Esta arquitectura presenta propiedades similares a la estrategia biológica como la robustez, estabilidad y aplicabilidad en un amplio rango de situaciones en los que la cooperación resulta ventajosa para el colectivo pero costosa para el individuo.

2. Desarrollo de un mecanismo de aprendizaje por refuerzo para sistemas multiagente que requieren altruismo.

   Se propone un algoritmo de aprendizaje por refuerzo que cumple las especificaciones de las reglas de aprendizaje locales en la naturaleza, en particular cumple con la Suma Relativa de Beneficios (RPS) presentando un balance óptimo en el dilema exploración-explotación. Este algoritmo utiliza información exclusivamente local a cada agente restringiéndose, de esta manera, los requerimientos de comunicación necesarios para realizar evaluaciones globales del rendimiento colectivo.

3. Delimitación del modelo de cooperación.

   Se presenta una caracterización del altruismo en el contexto de los sistemas multiagente estableciéndose de forma empírica los límites de las condiciones de aplicabilidad de la reciprocidad.

4. Implementación física de la arquitectura AREA en el robot COOBOT

   Se ha trasladado la arquitectura AREA al robot autónomo móvil (COOBOT) desarrollado en el departamento de Matemática Aplicada (Biomatemática). Este robot incluye los mismos esquemas de comportamiento que los agentes virtuales confirmando la validez de los mismos para el trabajo en entornos reales.

1.4 Planteamiento experimental.

Para el desarrollo de los experimentos presentados en esta memoria, se ha planteado un sistema multiagente realizando una tarea de recogida de objetos. Esta es una tarea frecuentemente utilizada en el contexto de robótica colectiva (Deneubourgh, 1991; Mataric, 1994b; 1996; Murciano y Millán, 1996; Murciano, Millán y Zamora, 1997).

La elección de este problema viene motivada por varias razones fundamentales. La primera de ellas es su clara inspiración en la tarea de forrajeo encontrada en los seres vivos. El forrajeo es uno de los fundamentales aspectos que deben solucionar los seres vivos y para ello han desarrollado multitud de estrategias con el fin de optimizar su eficacia. Existen muchos estudios de estrategias de forrajeo encontradas en los seres vivos. De entre ellos, los realizados bajo los auspicios de la Teoría de forrajeo óptimo (Pike *et al.*, 1977; Stephen y Krebs, 1986) asumen un conocimiento global del entorno de trabajo y suponen cierta racionalidad a los animales para elegir las acciones más indicadas en cada instante. Sin embargo, en la resolución de esta tarea mediante el sistema multiagente que se propone, se tiene que el conocimiento del mundo es parcial, la elección de la estrategia más adecuada se realiza exclusivamente con información local a los agentes y, como hemos visto, se presentan problemas de estabilidad. Estas circunstancias hacen inadecuado el estudio de este problema desde una perspectiva exclusivamente optimizadora como la que posee la teoría de optimalidad.

Otras razones que hacen atractiva la elección de esta tarea:

1. El colectivo obtiene beneficios si los individuos se comportan de forma cooperativa.

   El objetivo es que los agentes maximicen la cantidad de objetos recogidos del ambiente. La mejor estrategia posible a la vista de un objeto es recogerlo y transportarlo al lugar de almacenamiento en el menor tiempo posible. Ahora bien, si los objetos están agrupados en zonas poco accesibles, la mejor estrategia colectiva es que los agentes se detengan a señalizar al resto la presencia de objetos. La inclusión de comportamientos cooperativos de señalización de objetos incrementa la eficacia del colectivo (Murciano y Millán, 1996). Sin embargo, desde la perspectiva individual del señalizador, este comportamiento es poco eficiente y cumple con la definición de altruismo.

2. La eficacia de los distintos comportamientos depende del ambiente.

   Si la distribución de objetos cambia y éstos aparecen uniformemente distribuidos en el mundo de trabajo, en lugar de agrupados en zonas poco accesibles, la cooperación dejará de ser ventajosa. No tiene sentido en este caso, incurrir en pérdidas de tiempo de señalización dado que las ganancias que el colectivo obtendrá fruto de este comportamiento serán menores que el coste que supone para el sistema el tener un número de agentes detenidos señalizando en lugar de recolectando.

1.4 Esquema de la exposición.

La presente memoria esta estructurada en los siguientes capítulos:

Capítulo 1: Presenta las motivaciones y marco conceptual de la tesis así como su enunciado y las principales aportaciones.

Capítulo 2: Presenta los fundamentos que soportan las elecciones realizadas así como una breve descripción del concepto de agente autónomo. Discute el paralelismo entre la biología y la robótica en criterios de optimalidad y adaptabilidad. Discute sobre la cooperación en sistemas multiagente, sus problemas de estabilidad y las soluciones encontradas en la naturaleza y en los sistemas artificiales.

Capítulo 3: Presenta el estado del arte en robótica colectiva en relación con los modelos de cooperación así como una revisión de los trabajos relacionados con la estabilidad del altruismo.

Capítulo 4: Describe las condiciones experimentales y presenta la descripción de la arquitectura de reciprocidad propuesta (AREA). Presenta el algoritmo de aprendizaje utilizado así como las medidas elegidas para la evaluación del mismo.

Capítulo 5: Expone los resultados de la fase experimental donde se demuestra el problema de la estabilidad de la cooperación, la solución mediante la inclusión de la estrategia de reciprocidad así como la flexibilidad del algoritmo de aprendizaje para superar cambios ambientales. Muestra la escalabilidad del sistema propuesto y desarrolla un caso particular cuando la cooperación coexiste con la competencia.

Capítulo 6: Presenta los resultados de distintos experimentos que estudian los límites de la estrategia de reciprocidad. Discute las condiciones y el margen de aplicabilidad del modelo de cooperación propuesto.

Capítulo 7: Describe el robot autónomo COOBOT y presenta la evaluación de los comportamientos de la arquitectura AREA en esa plataforma. Presenta los resultados experimentales de la recogida de objetos por dos robots COOBOT discutiendo la adecuación de la arquitectura de reciprocidad para su implementación en robots físicos.

Capítulo 8: Presenta las conclusiones fundamentales de la tesis así como los trabajos futuros.

Capítulo 2

Fundamentos

Este capítulo se dedica a la exposición de los fundamentos que soportan las elecciones realizadas en la presente tesis. Comienza con una breve introducción al concepto de agente autónomo y una enumeración de sus propiedades, discutiendo los conceptos de optimalidad y adaptabilidad en los agentes autónomos, tanto naturales como artificiales. Discute sobre la cooperación como estrategia beneficiosa para incrementar el rendimiento de un colectivo de agentes, presentando los problemas de estabilidad de la misma. Revisa brevemente las diferentes teorías propuestas para explicar la evolución de la cooperación y en particular la Teoría de la Reciprocidad. El capítulo termina con una revisión del concepto de aprendizaje aplicado a sociedades de agentes autónomos, concretándolo en el mecanismo de aprendizaje por refuerzo.

2.1 Agentes autónomos.

El trabajo experimental de esta tesis se ha realizado sobre lo que se conoce como *agentes autónomos*. Se ha demostrado que el uso de estos agentes en la generación de sistemas

inteligentes es adecuado para afrontar problemas *situados* (Maes, 1991), como el abordado por esta tesis.

Estos problemas reúnen una serie de características que dificultan la búsqueda de soluciones eficientes basadas en razonamientos simbólicos sobre modelos del problema.

1. El dominio del problema es dinámico.

   Dado que el agente está "situado" en el entorno y sus acciones pueden modificarlo, éste se convierte en un ambiente extremadamente dinámico. Adicionalmente se tiene que el resultado de las interacciones agente-entorno son probabilísticas más que determinísticas y frecuentemente las leyes de probabilidad que las rigen varían. Esta variabilidad de las características ambientales y la impredecibilidad del resultado de las acciones del agente, dificultan los procesos de modelización global para su uso en sistemas de planificación.

2. Problemas muy complejos

   Frecuentemente se deben afrontar muchos sub-problemas o sub-tareas simultáneamente y éstos pueden variar en el tiempo. Esto obligaría a los sistemas planificadores a un replanteamiento del problema con el consiguiente incremento de requerimientos computacionales (tiempo, memoria, etc.).

3. Información incompleta y ruidosa.

   Otra característica de la "situación" del agente es que la información del dominio del problema la percibe directamente por sus sensores. La elección de una acción debe realizarse con este conocimiento que muchas veces es parcial y frecuentemente con ruido.

La necesidad de generar procesos inteligentes para solucionar problemas en estas circunstancias ha promovido el uso de los agentes autónomos. El desarrollo del concepto de agente autónomo está inspirado en gran medida en el estudio de sus homólogos naturales, los animales. Es por ello frecuente referirse a estos agentes con el término *animats* como abreviatura de "animal artificial" (Wilson, 1985).

Por *agente* se conoce a aquel sistema que trata de cumplir una serie de objetivos en una ambiente dinámico y complejo. Brooks (1991b) completa esta definición con la característica "irrenunciable" que deben poseer estos agentes como es la de poseer un cuerpo físico en un entorno real. Así, en una definición más estricta de agente tenemos que es un ente físico, generalmente móvil, que cuenta con sensores para percibir el mundo y con efectores (motores) para interaccionar con él. Maes (1994) relaja este requerimiento para dar cabida en esta definición a los agentes virtuales que trabajan en entornos también virtuales. Entre estos últimos se cuentan los agentes software para la recuperación de

información, para la formación interactiva, para la personalización de interfaces, etc. Se dice que un agente es *autónomo* si su comportamiento se rige con independencia de control externo, es decir, cuando toma sus propias decisiones a partir de la información que posee, tanto de su estado interno como de su entorno local. En el extremo opuesto a la autonomía se encuentran los denominados *autómatas,* los cuales carecen de cualquier autonomía y ejecutan invariablemente el programa para el cual han sido diseñados. Una barrera de un aparcamiento o una máquina de lavar coches son ejemplos de autómatas[1].

Un aspecto fundamental del trabajo con agentes autónomos es el diseño de su arquitectura de control. Este diseño debe contemplar la definición de los objetivos del agente, la definición de la información sensorial relevante, interna (en forma de estados motivacionales) y externa, y por último, la definición de la forma de integrar esta información sensorial para el desencadenamiento de una acción. Existen distintas aproximaciones al diseño de estos sistemas de control. La frontera entre las distintas arquitecturas que se han propuesto es difusa pudiendo clasificarse de acuerdo a varios criterios. Los fundamentales se basan en la existencia o no de procesos de planificación o deliberación y si las acciones se agrupan o no en comportamientos. Según estos criterios podemos distinguir dos tipos fundamentales de sistemas[2].

1. *Sistemas planificadores o deliberativos.*

   Estos sistemas usan modelos centralizados del mundo para verificar las lecturas sensoriales y para generar acciones. Esta arquitecturas tienen como problemas fundamentales los requerimientos computacionales (tiempo, memoria), poca adecuación a ambientes cambiantes y poca resistencia al ruido y la impredecibilidad de sensores y efectores.

2. *Sistemas reactivos*

   En los sistemas reactivos puros, entre percepción y acción no media planificación alguna. La acción motora se desencadena a partir de la actual lectura sensorial. Se consiguen funcionamientos en tiempo real pero resisten mal a la percepción errónea o incompleta del ambiente.

En el diseño de las arquitecturas de control, es frecuente recurrir a la definición de *comportamientos* como unidades básicas de construcción de la arquitectura. En un comportamiento se agrupa, en forma de pares, la información sensorial y las respuestas motoras de bajo nivel necesarias para una tarea concreta (por ejemplo, evitación de obstáculos). La información sensorial puede activar numerosos comportamientos de forma simultánea. Mediante listas de prioridades preprogramadas, sistemas de competición entre

[1] Para una revisión de los conceptos de autonomía y autosuficiencia ver por ejemplo McFarland (1994) o McFarland y Boesser (1993).

[2] Entre los dos extremos que se presentan, existen un gran número de abordajes híbridos que combinan cualidades de ambos.

comportamientos del tipo *winner-take-all* o normalización de comportamientos se determina la secuencia espacio-temporal de ejecución de acciones. Entre los sistemas *basados en comportamientos* han resultado de especial relevancia arquitecturas como las *arquitecturas subsumidas* (Brooks, 1986; 1990a; Maes, 1991) y *arquitecturas basadas en esquemas* (Agre y Chapman, 1987; Arkin, 1989; Balch y Arkin, 1994).

Las arquitecturas más aceptadas y de uso más frecuente en la actualidad, en el entorno de los sistemas multiagente descentralizados, combinan comportamientos reactivos y aprendizaje (Mataric, 1997; Murciano y Millán, 1996; Steels, 1996). Las propiedades más atractivas de estas arquitecturas son:

1. Pocos requerimientos computacionales y de memoria.

   No existen modelos globales del mundo de trabajo que consumirían grandes recursos de cálculo, tiempo y almacenamiento. Las acciones se deciden de forma distribuida en la arquitectura del agente a partir de información localmente obtenida. En la práctica, existen compiladores de arquitecturas (como el que ha desarrollado el MIT para las arquitecturas subsumidas (Brooks, 1990b)) que producen código ensamblador implementable en casi cualquier procesador de reducida capacidad (por ejemplo, en un Motorola 68HC11).

2. Gran velocidad de respuesta.

   Su estructura, compuesta de comportamientos básicos, funciona en paralelo. La existencia de esquemas reactivos acelera el proceso de decisión de acciones pues sólo requiere la búsqueda del par situación-acción adecuado y ejecutarlo. Las conexiones entre los distintos módulos no utilizan complejos procesos de deliberación sino que se limitan activarse o inhibirse con el consiguiente incremento en la velocidad de proceso.

3. Robustez.

   Resisten bien la existencia de ruido en la señal de sus sensores. La degradación que sufre el rendimiento del agente a medida que fallan sus sistemas sensoriales es gradual y paralelo al grado de malfuncionamiento. No existen inhabilitaciones completas del sistema por un fallo en alguno de sus sensores. Otra aportación a la robustez viene del hecho de que el comportamiento global está descompuesto en unidades básicas. No hay ningún módulo más importante que otro con la consiguiente ausencia de puntos críticos en el funcionamiento. Finalmente se pueden diseñar módulos redundantes de comportamiento para disminuir la posibilidad de errores.

4. Adaptabilidad.

   La inclusión en los agentes de mecanismos para aprender de la experiencia les capacitan para realizar cambios en sus comportamientos y así incrementar su

rendimiento. También se pueden incluir procesos de autocalibración de sensores y efectores para adaptarse a posibles cambios ambientales o malfuncionamiento de estos dispositivos.

5. Emergencia de nuevas funcionalidades.

   Es una de las más atractivas propiedades de los agentes autónomos. De la interacción de comportamientos dentro del agente (o de la interacción agente-entorno) puede surgir un nuevo comportamiento que no estaba diseñado. Aunque es de gran interés, no deja de ser problemático este efecto sobre la predecibilidad del funcionamiento del agente.

6. Integración de los agentes autónomos en sociedades (sistemas multiagente)

   Como veremos más adelante en éste capítulo, los agentes autónomos pueden ser fácilmente integrados en sociedades. El efecto fundamental es, nuevamente, la emergencia de funcionalidades fruto de las interacciones entre agentes. Estas van desde comportamientos colectivos emergentes básicos como la dispersión o el agrupamiento (Mataric, 1994b) hasta fenómenos de competencia, cooperación y especialización.

Todas estas propiedades hacen de los agentes autónomos una alternativa idónea para el desarrollo del trabajo que se proponen en esta tesis. En particular, utilizaremos una arquitectura de agente autónomo que incluye una serie de esquemas reactivos de comportamiento. Algunos de ellos son fijos mientras que otros son adaptables de forma dependiente del entorno (ver capítulo 4). El objetivo de esta adaptación será la decisión de qué *comportamiento social* mostrar en cada momento en función de las características del ambiente.

2.2 El comportamiento bajo el criterio de Optimalidad.

En el contexto de los agentes autónomos (naturales y artificiales), la toma de decisiones se rige por el criterio de *optimalidad*. Esto es, se persigue que el agente autónomo tenga un rendimiento óptimo (o cercano a él) en el desarrollo de la tarea para la que ha sido diseñado. A veces es suficiente que el agente cumpla con la propiedad de la *adecuación* (Brooks, 1991b). Según ésta, el mecanismo de selección de acciones del agente debe operar de forma que le haga cumplir los objetivos marcados a largo plazo. Por ejemplo, un agente realizando una tarea de navegación presentará un comportamiento aceptable (y

adecuado) si consigue recargar sus baterías antes de consumirlas aunque su trayectoria hacia el objetivo no sea óptima.

En este proceso de selección de acciones, el agente realiza una evaluación de las distintas alternativas de comportamiento eligiendo aquellas que maximicen una función conocida como *función de utilidad*[3] (o bien minimicen alguna función de coste). La definición de las distintas funciones de utilidad es difícil pues depende del objetivo del agente, frecuentemente es una función multidimensional ya que los objetivos suelen estar divididos en varios sub-objetivos, es dependiente del ambiente donde se desenvuelve el agente y del estado interno del mismo y, por último, depende del horizonte de tiempo en el que se desenvuelve el agente. De su correcta definición dependerá el grado de optimalidad alcanzado. Así por ejemplo, en el caso de un sistema de navegación, la función a minimizar será el tiempo tardado en alcanzar el objetivo, aunque esta función del tiempo puede tener otras dimensiones como por ejemplo navegar de la forma más segura posible, o no consumir las baterías totalmente, también pudiendo variar éstas según distintos estados internos del agente y distintos horizontes de tiempo, etc.

Aparentemente, el objetivo fundamental de los seres vivos es más claro. En este caso, es la extensión de su genotipo. Para su consecución, deben adoptar soluciones a las diferentes presiones selectivas que actúan sobre diferentes niveles tanto de sus estructuras físicas y fisiológicas como de sus repertorios de comportamiento. El resultado que en la actualidad presentan las distintas formas de vida viene de un compromiso entre esas presiones selectivas. Así, el diseño de los seres vivos se rige por un principio que McFarland (1993) denomina el principio del *diseño óptimo* que plantea que un sistema (animado o inanimado) que es capaz de modificarse en sucesivas versiones tenderá a evolucionar hacia una configuración óptima para las distintas presiones selectivas.

En el contexto biológico, al igual que en el caso de los agentes artificiales, la función de utilidad es característica del ambiente y de la actividad que se desarrolle. Se puede definir como el nivel de riesgo instantáneo en que incurre un animal (y los beneficios reproductores que obtiene) en un estado interno particular, realizando una determinada actividad en un ambiente concreto (McFarland, 1977). Por ejemplo, en el caso de la actividad de forrajeo, se debe maximizar la cantidad de alimento obtenido por unidad de tiempo a la vez que se deben minimizar distintos costes como por ejemplo el coste de encontrarse con depredadores (supervivencia), el tiempo que deja desprotegida a su prole (descendencia), etc. Finalmente, el comportamiento mostrado por el individuo será aquel que presente un mejor balance entre costes y beneficios. La cuantificación final de este balance se realiza en unidades de eficacia Darwinista. En éste sentido, los seres vivos se consideran como "máquinas" de maximizar su eficacia Darwinista (Hammerstein y

[3] En la literatura se encuentran distintas formas de referirse a estas funciones en los más variados contextos: funciones objetivo, funciones de valor (McFarland, 1994), funciones de evaluación (Barto, Sutton y Waltkins, 1989), funciones de optimalidad y aptitud darwiniana (Kazelnick y Bernstein, 1994). Todas ellas se refieren a la existencia de una función multivariable que debe optimizarse.

Selten, 1993). Esta eficacia es la suma de dos componentes que por un lado miden el éxito reproductor de un individuo y por otro su probabilidad de supervivencia. El resultado final mide la habilidad de un determinado genotipo de extenderse en la población a lo largo de la evolución. Una definición más amplia de la eficacia Darwinista es la debida a Hamilton (1964) que se conoce como *eficacia inclusiva*. En el cómputo de ésta se toma en cuenta, no sólo la eficacia (éxito reproductor y valor de supervivencia) propia de un individuo, sino que se toma en cuenta además la eficacia de los individuos genéticamente relacionados, es decir, su parentela.

El proceso de maximización de las funciones de utilidad es un proceso dinámico en el que se optimiza una medida de retroalimentación que proviene del ambiente y se usa como guía para alcanzar el objetivo para el que se ha diseñado el agente. Ahora bien, el comportamiento óptimo para un individuo depende del ambiente[4] y éste no es un conjunto de características inmutable sino que puede variar en el tiempo y en el espacio pudiendo ser además desconocido. Es por tanto imposible determinar a priori una secuencia de comportamientos para un agente que sea óptima para cualquier ambiente, es decir, un agente en un ambiente cambiante no alcanzará rendimientos óptimos en ausencia de procesos de adaptación.

2.3 Procesos adaptativos.

La adaptación es un proceso dinámico, que emerge de las interacciones agente-entorno y conduce a los individuos a modificar su comportamiento dependiendo de las distintas presiones que impone el ambiente, incrementando su grado de optimalidad. Los procesos adaptativos exploran las características relevantes del ambiente, generan alternativas de comportamiento para hacer frente a las distintas presiones ambientales y evalúan esas alternativas seleccionando las más eficaces.

2.3.1 Características generales de los procesos adaptativos.

Existen tres procesos fundamentales que los mecanismos de adaptación deben realizar:

1. Generación de variaciones en los comportamientos.

[4] Entendemos ambiente en su acepción más universal. En el caso de comportamientos sociales, un agente considera al resto de los individuos de la sociedad como formando parte de su ambiente.

Debe existir algún mecanismo, más o menos estocástico, de generación de alternativas de comportamiento que explore el espacio de soluciones posibles. Ahora bien, debe existir un balance apropiado entre la exploración de nuevas soluciones y la explotación de las actuales. La importancia relativa de estas dos estrategias, explorar-explotar, depende del efecto conjunto de varios factores. Entre ellos, tenemos:

- El grado de optimalidad ya alcanzado por el proceso adaptativo. A mayor grado de optimalidad menor exploración,
- La tasa de cambio del ambiente. Ante ambientes muy poco variables, el grado de exploración debe ser menor.
- El horizonte de tiempo del agente. Ante horizontes de tiempo cortos, es preferible la explotación de las soluciones, aunque éstas no sean óptimas, a la exploración de nuevas posibilidades.
- Máximos locales. Dado que la función a optimizar puede no ser monótona, una exploración insuficiente puede producir estancamientos del proceso en soluciones no óptimas.

2. Evaluación de las distintas alternativas propuestas.

   La forma de evaluación está íntimamente ligada a la función de optimalidad. Los mecanismos de evaluación se encargan de asignar "unidades de optimalidad" a las distintas alternativas. Se debe realizar una comparación de los beneficios obtenidos al mostrar cada una de las alternativas de comportamiento. Estas comparaciones se pueden realizar de forma absoluta o de forma relativa a algún valor esperado, o relativa al valor acumulado hasta el momento o hacia algún promedio de beneficio obtenido.

3. Selección de las alternativas más ventajosas.

   Tras la comparación se debe seleccionar, de forma duradera, alguna de las alternativas, bien por sustitución de otras menos rentables, o bien incrementando la posibilidad de mostrar dicha alternativa en sucesivas ocasiones.

Finalmente, los mecanismos de adaptación deben cumplir con los siguientes requisitos:

1. *Las adaptaciones deben ser incrementales.* La adaptación debe proceder a la vez que el agente se desenvuelve en su entorno haciendo que se incremente su eficacia progresivamente.
2. *Las adaptaciones deben orientarse hacia los aspectos críticos del ambiente,* evitando otros menos relevantes o de menor importancia en la función de optimalidad. Bajo la asunción de que toda adaptación es costosa, inversiones en

adaptaciones a aspectos poco relevantes del ambiente distraen recursos para la adaptación a otros aspectos más importantes.

3. *El proceso de adaptación debe ser autónomo* (no supervisado). No es posible disponer de un maestro que proporcione al agente la asociación correcta entre una situación ambiental y la respuesta adecuada.

De entre los mecanismos de adaptación al medio en los seres vivos, podemos distinguir aquellos realizados en tiempo somático (*adaptaciones fenotípicas*) y aquellos realizados en tiempo evolutivo (*adaptaciones genotípicas*). En palabras de Bateson (1979):

> *"We face, then, two great stochastic systems that are partly in interacction and partly isolated from each other. One system is within the individual and is called learning; the other is immanent in heredity and in populations and is called evolution. One is a matter of single lifetime; the other is a matter of multiple generations of many individuals".*

2.3.2 Adaptaciones fenotípicas.

Estas adaptaciones se producen en el tiempo de vida del individuo o tiempo somático. En este apartado se incluyen tanto los procesos de aprendizaje, como los de maduración o modificaciones fisiológicas temporales[5]. El objetivo de todos ellos es modificar el comportamiento del individuo para conseguir, por un lado rendimientos óptimos (o cuasi-óptimos) y por otro mantener las variables de estado del individuo dentro de su espacio de viabilidad (Meyer y Guillot, 1990). Nos centraremos en la descripción de los procesos de aprendizaje en el contexto de la toma de decisiones, dado que la selección de comportamientos es el problema que se aborda en el desarrollo de la tesis y el aprendizaje el mecanismo de adaptación que hemos incluido en la arquitectura de los agentes autónomos.

El aprendizaje se define como la modificación en el comportamiento de un individuo, fruto de la interacción con el ambiente, que perdura a lo largo del tiempo. Esta modificación tiene un sustrato neuronal. De una forma reduccionista se puede hablar de que las conexiones sinápticas entre neuronas varían su eficacia, modificándose así la acción

[5] A nuestro entender, los procesos de adaptación fisiológica y los procesos de maduración pueden ser considerados como adaptaciones fenotípicas dado que se manifiestan en el tiempo de vida del individuo. Pero puesto que se limitan a responder de forma invariable y reactiva al estímulo desencadenante y los mecanismos subyacentes a esa respuesta son heredables y seleccionados evolutivamente, pueden ser consideradas también como adaptaciones genotípicas (ver más adelante). En cualquier caso, evitaremos las referencias a estas adaptaciones y nos centraremos en el aprendizaje como adaptación fenotípica fundamental.

desencadenada por una determinada estimulación. El aprendizaje puede ser no asociativo o asociativo[6]. En este último caso se pretende relacionar adecuadamente qué acciones (comportamientos) se deben mostrar ante qué situaciones (percepciones).

El mecanismo de aprendizaje asociativo se realiza mediante la estrategia de *ensayo y error.* Ante una situación percibida, el animal muestra un comportamiento. Éste es evaluado de distintas maneras, pero en general, lo que se mide es el beneficio inmediato obtenido de la ejecución de ese comportamiento ante esa situación. Si bien se está asumiendo la inmediatez de las consecuencias del comportamiento, es posible también tratar con situaciones en las que las consecuencias de un comportamiento no son inmediatas. Finalmente, si la evaluación es positiva se procede a las modificaciones sinápticas pertinentes para que ante esa misma situación, la probabilidad de volver a mostrar ese mismo comportamiento se vea incrementada. En el caso contrario, si la evaluación resulta negativa, se disminuye esa probabilidad. La iteración de este procedimiento garantiza que el individuo elegirá aquellas secuencias de comportamiento que maximicen el beneficio obtenido, o al menos, aquellas que hasta el momento han resultado en un mayor beneficio acumulado en relación con el total de beneficios obtenidos con cualquier otra alternativa. Excepción a esta patrón de selección de acciones es el caso de los comportamientos exploratorios que se muestran de forma aleatoria o por mecanismos de imitación.

Para una recopilación de ejemplos interesantes de comportamientos sujetos a aprendizaje en la naturaleza ver por ejemplo Alcock (1993), Colmenares y Gómez (1994) o McFarland (1993).

2.3.3 Adaptaciones genotípicas.

Estas adaptaciones tienen lugar, como se ha dicho, en escala de tiempo evolutivo. En esta escala, la unidad de tiempo la constituye el tiempo inter-generacional. El sustrato de modificación es, en este caso, el material genético. El proceso evolutivo consiste en la generación aleatoria de variaciones en el genotipo de los individuos por medio de mecanismos de mutación y recombinación en la reproducción. La evaluación de las distintas alternativas se efectúa de acuerdo a los criterios de eficacia inclusiva. Aquellos fenotipos, como realizaciones de genotipos, que resulten en una mejor evaluación, serán seleccionados por efecto de la selección natural. La iteración de este proceso lleva a un proceso de optimización de forma que los individuos de una especie irán progresivamente seleccionando repertorios de comportamiento más adecuados para el ambiente en el que se

[6] Existen otros tipos de aprendizaje relacionados con otros procesos como la generalización, abstracción, categorización, etc. de mayor complejidad que el que se presenta. Están menos relacionados con los mecanismos de selección de acciones y más con el procesamiento simbólico de la información. De ahí nuestra elección del mecanismo de ensayo-error para la exposición.

desenvuelven. Una vez seleccionados permanecerán en el repertorio de la especie heredándose de generación en generación.

Los comportamientos que se seleccionan mediante este mecanismo se conocen como *comportamientos de disparo inmediato,* dado que para su desencadenamiento no es necesaria la experiencia previa sino que la simple percepción de la situación desencadenante provoca el disparo de la acción. Ejemplos de este tipo de comportamientos son, por ejemplo, las estrategias de reproducción de los insectos, estrategias de cortejo, las estrategias defensivas de huida, el mimetismo Batesiano, los mecanismos de impronta, etc.

Además de la diferencia señalada entre ambos mecanismos, según la escala de tiempo en que se producen, existe otra diferencia fundamental. El alcance de la evaluación realizada por ambos mecanismos. En el caso de los mecanismos de aprendizaje somático, la evaluación tiene carácter local al individuo, es decir, sólo se computa el beneficio recibido por el individuo como consecuencia de su comportamiento. En las adaptaciones genotípicas, para el cómputo de la eficacia de una estrategia de comportamiento no sólo se tiene en cuenta el éxito individual -eficacia Darwinista- sino también la eficacia debida al éxito de los individuos genéticamente relacionados -eficacia inclusiva.

Finalmente, el proceso adaptativo es producto de la acción conjunta de ambos mecanismos. El aprendizaje evolutivo se encarga de disponer las estructuras anatómicas capaces de aprender y los comportamientos de disparo inmediato. La actuación posterior del aprendizaje, a partir de las experiencias obtenidas de la interacción con el ambiente, modificará la arquitectura heredada continuando así el proceso adaptativo[7].

2.3.4 Modelos de adaptación.

La Inteligencia Artificial ha dedicado grandes esfuerzos a la modelización de los procesos de adaptación, fenotípica y genotípica, para la síntesis de comportamientos inteligentes. Esto ha dado lugar a una serie de modelos que se pueden clasificar, según la inspiración de las técnicas usadas, en Modelos Conexionistas y Algoritmos Genéticos. Es importante notar, que aunque pueda existir ésta división, en realidad se trata de la modelización de el mismo proceso de ajuste al medio, pero que sucede en escalas de tiempo distintas, esto es, tiempo somático y tiempo evolutivo.

En una breve descripción de los algoritmos genéticos, diremos que son una técnica de optimización que usa los mecanismos de la evolución y la selección. La técnica en sí,

[7] La exposición se realiza bajo una óptica no lamarckiana que asume unidireccionalidad en la relación entre ambos mecanismos. Se ha descrito, sin embargo, la posible influencia que el aprendizaje, en determinadas circunstancias, puede ejercer sobre el proceso evolutivo. Este efecto se conoce como el *efecto Baldwin* (Ackley y Littman, 1991; Baldwin, 1896)

consiste en la generación de una población de cromosomas artificiales que codifican posibles soluciones a un problema de optimización. Tras procesos exploratorios como la mutación, recombinación o la delección, se eligen aquellos cromosomas que obtengan mejor rendimiento frente a una función de evaluación. Estos son utilizados como parentales en la producción de una nueva generación filial con la que se repite el proceso (Holland, 1992; Koza, 1992; Wagner y Altenberg, 1996). Estas técnicas se han utilizado con buenos resultados en el control de agentes autónomos (Clift, Harvey y Husbands, 1993; Floreano y Mondada, 1996; Nolfi, 1997; Wilson, 1985).

Los modelos conexionistas tienen como objetivo la modelización de un determinado proceso, inspirándose en las redes neuronales biológicas y aprovechando las características de procesamiento eminentemente paralelo de las mismas. Una red neuronal artificial consiste en un conjunto de elementos de proceso simples, interconectados entre sí. La información fluye por esa red en forma de valores de activación de los distintos elementos, ponderados por coeficientes (pesos sinápticos) que emulan la eficacia sináptica. Los valores de los pesos sinápticos se alteran mediante diversos algoritmos de aprendizaje para conseguir un mapeo correcto entre las entradas de la red y las salidas adecuadas. En función de la consideración que se tenga de esa salida adecuada, tendremos distintos tipos de algoritmos de aprendizaje. Para una revisión general del aprendizaje en redes neuronales artificiales ver (Hecht-Nielsen, 1990; Hinton, 1989; Torras, 1985).

Se habla de algoritmos supervisados, como el de *retropropagación de errores* (Rumelhart, Hinton y Williams, 1986) cuando la salida deseada es conocida y lo que se pretende minimizar es el error cometido en la salida obtenida.

Si no se tiene información acerca de la salida adecuada, sino lo que se pretende es que la matriz de pesos se organice de acuerdo a patrones internos de los entradas realizando así un agrupamiento de las mismas, se habla de algoritmos no supervisados. Un ejemplo de estos modelos de aprendizaje lo constituyen los *mapas autoorganizativos* de Kohonen (SOM) (Kohonen, 1988).

Finalmente, si la información disponible de las salidas de la red es únicamente una idea de su bondad o adecuación al cumplimiento de un objetivo, estaremos ante un algoritmo de refuerzo. Esta evaluación de la utilidad de la salida de la red es lo que se conoce como señal de refuerzo y se utiliza para modificar los pesos de la red. El objetivo es maximizar el valor de las señales de refuerzo obtenidas, o sea maximizar la utilidad de las salidas de la red. La regla de aprendizaje más utilizada dentro de este tipo es la *búsqueda asociativa* (Barto, Sutton y Brower, 1981).

Elegiremos este último caso, el aprendizaje por refuerzo, como modelo de aprendizaje en el desarrollo de la tesis pues es el más próximo al planteamiento experimental de la misma. En el proceso de selección de comportamientos por parte de un agente, no se tiene información precisa de cual es el comportamiento adecuado para una situación determinada (y aún teniéndola, dado el carácter autónomo del agente sería imposible el uso

de la misma) sino que mediante un procedimiento de ensayo y error se obtienen una medida, más o menos precisa, de la utilidad de un comportamiento. Por otro lado, los algoritmos de refuerzo son los que presentan una mayor cercanía al aprendizaje animal de los tipos del condicionamiento clásico, en cuanto a la asociación de estímulos, (Sutton y Barto, 1987) y condicionamiento instrumental en cuanto a la metodología utilizada (Kaelbling, Littman y Moore, 1996). Una visión del concepto de aprendizaje desde una perspectiva biológica puede obtenerse en Domjan (1993).

Es importante señalar que la evaluación de la utilidad de los comportamientos se puede materializar de distintas formas, algunas de ellas de marcado carácter biológico. Por ejemplo, pueden realizarse modelos de aprendizaje que evalúen las acciones realizadas utilizando el valor adaptativo como medida de eficacia. La serie de autómatas Darwin (III y IV) debidas a G. Edelman y su grupo (Edelman *et al.*, 1992; Reek, Sporns y Edelman, 1990; Reek *et al.*, 1990) utilizan el valor adaptativo como señal para el aprendizaje dentro del marco de la Teoría del Darwinismo Neuronal (Edelman, 1987). Otros ejemplos del uso del valor adaptativo, en el aprendizaje de procesos de foveación en el sistema oculomotor, pueden encontrarse en (Murciano, Zamora y Reviriego, 1993; Murciano y Zamora, 1993).

El algoritmo básico de aprendizaje por refuerzo se fundamenta en un proceso de ensayo y error. El agente aprende actuando y no requiere de ningún maestro que proponga acciones correctas para todas las posibles situaciones en las que el agente puede encontrarse. Por el contrario, el agente prueba una acción y percibe una señal de retroalimentación, denominada señal de refuerzo, que cuantifica la utilidad de la acción realizada. Si esta señal es positiva, el agente tenderá a mostrar el mismo comportamiento en el futuro, evitándolo en caso contrario. El esquema general del algoritmo de refuerzo es como sigue (Kaelbling, 1990):

1. Inicializar el estado interno del agente $S = S_0$
2. Repetir siempre:
   - 2.1 Observar el estado actual del entorno I.
   - 2.2 Elegir una acción $a = V(I, S)$ de acuerdo a la función de evaluación V.
   - 2.3 Ejecutar la acción a.
   - 2.4 Recibir la señal de refuerzo r correspondiente a la ejecución de a en el estado I.

2.5 Actualizar el estado interno del agente, $S_{t+1} = U\ (S_t, I, a, r)$ usando la función de actualización U

La aplicación de este aprendizaje se ha demostrado adecuada en el control de agentes autónomos, en tareas como la navegación (Millán, 1996), control dinámico de efectores (Gullapalli, 1995; Martín y Millán, 1997a; 1997b) y otra variedad de tareas (Hashimoto *et al.* 1992; Lin, 1992; Mahadevan y Connell, 1992; Millán, 1995; Millán y Torras, 1992; Mitchell y Thrun, 1993; Prescott y Mayhew, 1992; Sutton, 1990).

2.4 Colectivos de agentes autónomos: sistemas multiagente.

Si observamos las formas de organización de los seres vivos, es frecuente encontrar que éstos se agrupan en colectivos o sociedades. Estos grupos sociales pueden ser desde sociedades de animales superiores, grupos familiares, grupos estables de nidificación o de caza, hasta sociedades de insectos eusociales, consideradas por algunos autores como las verdaderas sociedades (Hölldobler y Wilson, 1994).

Los comportamientos sociales confieren beneficios a los miembros del grupo (Alexander, 1974). El caso paradigmático de estas ventajas de la colectividad lo protagonizan las hormigas, que mediante una compleja organización social han conseguido un rotundo éxito evolutivo consiguiendo colonizar los más variados hábitats. Otros ejemplos de los beneficios de la socialidad son, por ejemplo, la reducción en la presión de los depredadores mediante comportamientos de alarma o de defensa colectiva, mejoras en la eficacia de la búsqueda de recursos y en la defensa de los mismos ante otros grupos, incremento en la eficacia de cuidado de la prole, etc. Sin embargo, la socialidad también conlleva una serie de costes. Entre ellos, aparecen un mayor número de interferencias entre animales del mismo grupo, que competirán por el espacio, la reproducción, el alimento, e incluso existirán mayores riesgos de que individuos dentro del grupo exploten al resto en sus esfuerzos cooperativos.

Estas ventajas, sin embargo, dependen de las condiciones ambientales. Por ejemplo, se observa una gran relación entre la distribución y accesibilidad de los recursos y la existencia de distintos grados de estructura social en los animales. Esta gradación, en el contexto de búsqueda de alimento, influye en el tamaño del grupo y en el solapamiento o no de las zonas de forrajeo. Así, de acuerdo al principio de Horn sobre forrajeo en grupo se

tiene que si un recurso está uniformemente distribuido, será mejor particionarse el territorio individualmente que formar colonias forrajeando colectivamente. Observaciones sobre el terreno, en grupos de aves, corroboran esta afirmación (Wilson, 1975).

La existencia de una relación coste/beneficio en los comportamientos sociales y la influencia que el ambiente tiene sobre dicha relación, nos lleva a seguir un razonamiento similar al llevado cuando se trataba de optimalidad individual en los agentes autónomos. Este es, en resumen el hilo argumental: si un comportamiento social (por ejemplo la cooperación) es ventajoso para el individuo, éste deberá ser seleccionado por medio de los mecanismos adaptativos mencionados en el apartado anterior. Si el ambiente cambia y se modificara el sentido de esa relación, es decir, si el comportamiento social dejara de ser ventajoso, los mecanismos adaptativos deberán garantizar que el agente seleccione otra alternativa para ese comportamiento. Veremos, en la siguiente sección, que este planteamiento se modifica en cierta medida debido a la aparición de problemas de estabilidad inherentes a los comportamientos altruistas.

La Inteligencia Artificial, como no, se ha inspirado en el estudio de los colectivos animales, tratando de extraer los comportamientos presentes en dichas sociedades para trasladarlos al entorno artificial. Fruto de este flujo de conocimientos, desde la observación de las sociedades biológicas hacia la robótica inteligente, en la última década ha surgido un nuevo campo de investigación denominado robótica colectiva (también sistemas multiagente). En suma, se trata de la creación de equipos de trabajo formados por colecciones de agentes autónomos. Además de las propiedades inherentes a dichos agentes, de las interacciones entre ellos emergen nuevas propiedades. Estas son fundamentalmente:

1. Mayor robustez del sistema.

   Dada la redundancia intrínseca del sistema, el mal funcionamiento de alguno de los agentes no paraliza el cumplimiento final de la tarea sino que provoca una degradación gradual del rendimiento.

2. Emergencia de nuevos comportamientos.

   Al igual que la interacción de los diversos comportamientos individuales dentro de un agente provocan la emergencia de funcionalidades nuevas (sección 2.1), las interacciones entre agentes provocan la aparición de comportamientos sociales más complejos. Por ejemplo, del comportamiento individual de evitación de obstáculos puede emerger el comportamiento colectivo de dispersión. Mataric (1994b) proporciona una interesante colección de este tipo de comportamientos colectivos emergentes.

3. Incremento en el rendimiento del sistema.

   3.1 La inclusión de estrategias de cooperación entre agentes abre el camino a incrementos en el rendimiento global del sistema. Desde las situaciones en las

que el rendimiento del sistema se incremente linealmente (un agente durante un tiempo t realice la misma cantidad de trabajo que t agentes en una unidad de tiempo) hasta otras en las que el efecto del incremento de agentes no sea de carácter sinérgico. (ver, por ejemplo, Murciano y Millán, 1996)

3.2 Las estrategias de especialización pueden incrementar el rendimiento del colectivo vía una mejor distribución de los recursos en distintas subtareas. En (Murciano, Millán y Zamora, 1997) se presenta un ejemplo de este tipo.

3.3 El funcionamiento paralelo de los sistemas multiagente incrementa también el rendimiento del sistema en aquellos problemas que puedan dividirse en subtareas.

4. Nuevas competencias.

   La posibilidad de realizar trabajos cooperativos amplia el espectro de tareas que pueden realizar los agentes. Muchos tareas son inherentemente distribuidas, en el espacio o en el tiempo o en sus partes, siendo por tanto necesario resoverlas de forma distribuida. En otros casos, las tareas requieren un número determinado de agentes para llevarlas a cabo, como por ejemplo el transporte de un objeto muy pesado que puede necesitar la cooperación de 2 o más agentes.

Además de las ventajas mencionadas, el empleo de sistemas multiagente plantea una serie de dificultades relacionadas fundamentalmente con los procesos de control y de evaluación del sistema.

1. Dificultades en la coordinación.

   Al integrar un número de agentes en un colectivo trabajando simultáneamente, se producen una serie de efectos negativos que dificultan la coordinación del sistema. Estos son, fundamentalmente:

   1.1 Incremento de las interferencias entre agentes (colisiones, etc.).

   1.2 Efectos de competencia en la realización de funciones (fruto de distribuciones no óptimas de los recursos, etc.).

   1.3 Aumento de la complejidad. Al aumentar el número de agentes, el espacio de soluciones crece exponencialmente al igual que lo hace el espacio de estados del sistema.

2. Dificultades en la evaluación.

La aparición de propiedades emergentes a partir de las interacciones de los agentes disminuye la previsibilidad de los resultados del trabajo dificultando la evaluación del mismo.

Debe hacerse notar que los efectos de estos problemas dependen en gran medida de la estrategia de control utilizada. Las alternativas para llevar a cabo este control se diferencian en la existencia o no de un controlador central del sistema. Se denomina control centralizado cuando existe un controlador que posee información de la tarea a realizar, el estado del ambiente y de todos los agentes, utilizando esta información para la coordinación del equipo de trabajo. Este tipo de control tiene numerosos puntos críticos, fundamentalmente causados por los requerimientos de comunicación y por su sensibilidad a los incrementos en la complejidad del sistema. Una alternativa atractiva es el control descentralizado, en el que no hay un controlador central sino que el control emerge de las interacciones entre los agentes. No es necesaria la existencia de modelos del mundo ni planificaciones, sino que los agentes usan la información percibida por sus sensores para controlar sus propios comportamientos, entre ellos, los comportamientos sociales, que posibilitarán la coordinación del sistema.

Finalmente, como en el caso de los agentes autónomos individuales, el trabajo con sistemas multiagente persigue los objetivos de optimalidad y adaptabilidad. La inclusión de mecanismos de cooperación entre agentes puede facilitar la consecución de rendimientos óptimos en el sistema. Sin embargo, cuando estos comportamientos cooperativos son altruistas (el caso extremo de la cooperación) y los agentes pueden adaptar sus comportamientos para maximizar su rendimiento individual, aparece un grave problema: la estabilidad del altruismo. El siguiente apartado se dedica al enunciado de este problema así como a la descripción de las estrategias observadas en la naturaleza que lo solucionan. Del estudio de las mismas, y del conocimiento de las restricciones que el trabajo con sistemas multiagente impone, se sugerirá una estrategia para estabilizar el altruismo en la robótica colectiva.

2.5 Evolución de la Cooperación: inestabilidad del altruismo.

Entre los comportamientos sociales, el estudio de la cooperación es el que sin duda ha despertado mayor interés en la comunidad científica, en la biología en particular aunque también en otras parcelas del saber como la economía y la sociología (Axelrod, 1984). Ya

a principios de este siglo, el ideólogo anarquista Kropotkin postulaba en su libro *El apoyo mutuo, un factor de la evolución,* que la cooperación humana podría emerger en ausencia de cualquier autoridad estatal que gobernara las interacciones entre individuos. La ubicuidad de la cooperación, incluso en condiciones de irracionalidad y completa autonomía, es una de las razones que justifican el interés despertado. Sin embargo, desde una perspectiva biológica, lo más llamativo de la cooperación es su aparente contraposición a la Teoría de la Evolución. Ésta se basa en la lucha por la supervivencia y en la selección natural de los más adaptados. Comportamientos que redujeran la eficacia relativa de los actores estarán en desventaja y tenderán a desaparecer. Esto ocurre con los comportamientos altruistas, los cuales provocan un coste para el actor de los mismos, siendo los receptores de dichos actos altruistas beneficiados en esa interacción (Tabla 2.1). Según este razonamiento, nada impediría que un individuo adoptando una estrategia no social -egoísta-, dentro de un colectivo de individuos altruistas, se extendiera en la población debido a su mayor éxito relativo en términos de eficacia. Esto es, incrementaría su eficacia como receptor de los actos altruistas sin pagar ningún coste. Planteemos un ejemplo: una de las tareas en las que se invierte un porcentaje del tiempo muy alto en muchas especies es la vigilancia frente a los depredadores. En ciertas especies, ante la presencia de un depredador, los individuos muestran un comportamiento de emisión de señales de alarma hacia el resto del grupo. Este comportamiento, aunque beneficioso para el resto de los miembros del grupo, repercute negativamente en el individuo que lo muestra, pues además de provocar pérdidas de tiempo a la hora de escapar, hace que el señalizador incurra en mayores riesgos al llamar la atención del depredador hacia sí mismo. Si en el seno de ese grupo apareciese un mutante para ese comportamiento de forma que tuviera una actitud *egoísta*, obtendría las ventajas de usar las señales de alarma de sus compañeros sin incurrir nunca en el coste de señalizar. En ausencia de autoridad alguna que evite este tipo de comportamientos abusivos[8], nada impediría que bajo las leyes de la selección natural, estos individuos egoístas extendieran su genotipo por la población desplazando al carácter altruista de la misma.

Adicionalmente a este efecto de invasión del egoísmo en una población de altruistas, se tiene otro efecto paradójico. La ventaja que obtiene un individuo mutante egoísta depende de la presencia o no de individuos altruistas en la población. Siguiendo con el ejemplo que hemos planteado, es fácil comprobar que la eficacia de un individuo egoísta en un grupo de animales que están forrajeando mientras vigilan la presencia de depredadores, depende de la existencia de individuos señalizadores altruistas dentro de ese grupo. Si desaparece el carácter altruista de la población al ser completamente invadida por el carácter egoísta, la eficacia individual de los egoístas sería, en promedio, inferior incluso que la que

[8] Recientemente se ha demostrado que en grupos sociales de primates existen acciones de castigo hacia individuos no-sociales que no cooperan con el resto del grupo. En estos casos los individuos dominantes ejercen su autoridad evitando que haya individuos que abusen de los esfuerzos cooperadores del resto (Clutton-Brock y Parker, 1995).

presentaban cuando se comportaban altruistamente asumiendo los costes de la señalización.

Esta tendencia de ciertas estrategias de comportamiento a ser desplazadas del patrimonio genético de las poblaciones, debido a la invasión por otras estrategias mutantes más eficaces, es lo que se conoce como inestabilidad evolutiva. La definición del concepto de *estabilidad evolutiva* se debe a Maynard-Smith y Price (1973).

Formalmente, sea $E(I,J)$ la eficacia de un individuo con comportamiento I dentro de una población de individuos con comportamiento J, y sea $E(J,P_{q,J,I})$ la eficacia de una comportamiento J en una población P formada por q individuos con comportamiento J y $(1-q)$ individuos I. La estrategia I será una *estrategia evolutivamente estable* (ESS) bajo las leyes de la selección natural si, para todo $J \neq I$, se cumple alguna de las siguientes expresiones (Maynard-Smith, 1982):

$$E(I,I) > E(J,I)$$

o bien

$$E(I,I) = E(J,I) \text{ y para una pequeña proporción } q$$
$$E(I,\ P_{q,J,I}) > E(J,\ P_{q,J,I})$$

En el caso del comportamiento de señales de alarma ante depredadores, se tiene que no es una ESS dado que la eficacia de un egoísta dentro de un colectivo de altruistas es superior a la de un altruista en ese mismo colectivo. Consecuencia de ello es su susceptibilidad de invasión por la estrategia egoísta. A pesar de la inestabilidad manifiesta de las estrategias altruistas, éstas están presentes en un gran número de sociedades animales.

Los sistemas multiagente con aprendizaje, en los que los agentes deben maximizar su rendimiento individual, se enfrentan con el mismo problema de inestabilidad del altruismo. En ese sentido paralelizan fielmente lo que sucede en la naturaleza. Imaginemos un sistema de esas características, es decir, con aprendizaje, en el que además los agentes pueden mostrar un comportamiento cooperativo de tipo altruista. Este comportamiento, por definición, supone una pérdida de eficacia para el actor altruista mientras que el receptor del mismo resulta beneficiado. Si los agentes pueden modificar su comportamiento mediante un mecanismo de aprendizaje por refuerzo que tienda a seleccionar aquellos comportamientos más ventajosos para el individuo, nada impediría que los individuos seleccionasen la alternativa más rentable, esto es, el comportamiento egoísta, abandonando la más costosa, la altruista.

Tabla 2.1. Cambios en la eficacia de actores y receptores de algunas de las posibles interacciones entre individuos. Un signo menos significa una pérdida de eficacia mientras que un signo más significa una ganancia. Las flechas indican el actor original del acto y el receptor del mismo. Se asume independencia genética entre actores y receptores.

Tipo de interacción	Cambios en la eficacia		
Mutualismo	+	⟶	+
Altruismo	−	⟶	+
Comportamiento egoísta	+	⟶	−
Comportamiento de castigo	−	⟶	−−
Reciprocidad	−	⟶	++
	+ +	⟵	−

Es importante resaltar que el problema de la estabilidad del altruismo surge de la coexistencia, en una misma arquitectura, de un mecanismo de aprendizaje y la existencia real de coste en la cooperación que hace que ésta sea realmente altruista. Tanto la adaptación como el altruismo son propiedades de gran interés para su inclusión en los sistemas multiagente.

Una vez más nos encontramos con un paralelismo entre la problemática presente en la naturaleza y los sistemas robóticos colectivos. Las soluciones que se presenten en los sistemas biológicos para la estabilización del altruismo podrán inspirar modelos de cooperación válidos para su traslado a los sistemas multiagente. En la siguiente sección se describen estas soluciones así como los escasos equivalentes en los sistemas multiagente artificiales existentes en la actualidad.

2.6 Mecanismos de estabilización del altruismo.

La presencia, aparentemente paradójica bajo la óptica de la teoría de la evolución, de comportamientos altruistas en los colectivos animales ha llamado la atención a numerosos científicos durante las últimas décadas. A continuación haremos una breve exposición de las teorías surgidas para explicar esta cuestión.

En los años 60, Wynne-Edwards (1962) propuso el argumento del *bien de la especies* como justificación de la presencia de comportamientos que, aparentemente, estaban en contraposición con los principios de la selección natural. Así nacía la teoría de la *Selección de grupo.* Según ésta, las especies que contaran con mecanismos de autorregulación de los tamaños de sus poblaciones así como de las tasas de consumo de sus recursos, sobrevivirían. Por el contrario, aquellas que carecieran de dichos mecanismos se extinguirían por sobre-explotación de los recursos. De esta forma sería posible explicar, por ejemplo, el fenómeno de aparente suicidio colectivo que presenta una especie de mamíferos, los "lemmings" (*Lemmus lemmus*). En esta especie, cuando la densidad de la población alcanza un punto crítico, se desencadena un comportamiento migratorio que en ocasiones conduce a un autosacrificio por el que se regula el tamaño de la población. Se evitará así la sobre-explotación de los recursos y la desaparición de la especie. Según esta teoría, la selección natural operaría a nivel de grupo en lugar de a nivel individual, beneficiando y seleccionando aquellos grupos con comportamientos ventajosos para el colectivo. El principal argumento en contra de ésta teoría es que la selección natural tiene mucha más fuerza a la hora de configurar la composición genética de posteriores generaciones si actúa sobre las variantes genotípicas individuales en lugar de cuando actúa sobre las variaciones de los grupos (Williams, 1966; Krebs y Davies, 1993). Aunque todavía existe cierta controversia en torno a la teoría de selección de grupo (Wilson, 1979; Wynne-Edwards, 1986), está aceptada la visión más próxima a la original de Darwin que enfatiza el carácter individual de la selección natural.

De acuerdo a esta visión más ortodoxa de selección natural sobre los individuos, se han propuesto diversas alternativas para explicar la aparición y posterior estabilización de comportamientos altruistas. El fundamento de todas ellas consiste en la compensación del coste de mostrar comportamientos altruistas mediante el incremento de la eficacia inclusiva del actor altruista en, al menos, una de las siguientes vías:

i) Incrementar la eficacia de los individuos relacionados genéticamente.

ii) Incrementar la eficacia Darwinista individual.

En ambos casos se debe conseguir una ganancia neta superior al coste invertido en el acto altruista. De no ser así, los actores de estrategias altruistas estarán en desventaja frente a los receptores de dichos comportamientos siendo pues inestables. Si bien en esta

exposición se enuncian como mecanismos separados, su acción en la naturaleza es conjunta. La importancia relativa de cada uno de ellos en el establecimiento de un determinado comportamiento altruista es de difícil cuantificación. Esta división de los mecanismos concuerda, en casi todos sus términos, con la división que puede hacerse de los comportamientos altruistas según sean mostrados hacia individuos genéticamente relacionados (apartado i), o hacia individuos no relacionados genéticamente (ii).

La vía de compensación de los costes de actos altruistas mediante el incremento de la eficacia de los individuos genéticamente relacionados, origina la que se conoce como la *Teoría de Selección por Parentesco* (Maynard-Smith, 1964). Según ésta, el coste del acto altruista se ve compensado por la ganancia neta en número de copias de genes del individuo que pasan a la generación siguiente. Es decir, dado que las formas de reproducción constituyen una transmisión de patrimonio genético, los individuos genéticamente relacionados compartirán patrimonio genético en un grado proporcional a su relación de parentesco. En especies diploides con reproducción sexual, la probabilidad de que un hijo porte un gen de sus parentales es ½, mientras que en el caso de sobrinos y nietos es ¼ y en bisnietos y primos es $^{1}/_{8}$. Estas probabilidades se denominan grado de parentesco. Por lo tanto, un acto altruista extremo, como sería el caso del sacrificio de la vida de un padre para proteger a su descendencia, supondría la pérdida de una copia de ese gen, pero ésta inversión habrá sido rentable en términos genéticos si ha supuesto salvar la vida, en términos promedio, de más de dos hijos, más de cuatro sobrinos o nietos o más de ocho primos o bisnietos. Paralelo al mecanismo de selección por parentesco, debe existir un mecanismo fiable de reconocimiento del mismo para seleccionar correctamente a los destinatarios de los actos altruistas (Herper, 1991; Fletcher y Michener, 1987). Una de las soluciones más comunes para este reconocimiento consiste en la asunción de que los individuos emparentados viven cercanos entre sí. De ésta forma, mediante la impronta es posible asociar, durante las primeras experiencias de vida, a los individuos del entorno familiar. Se ha propuesto también la existencia de alelos de reconocimiento como forma de implementar este discriminación de parentesco (Hamilton, 1964).

La naturaleza ofrece muchos ejemplos de estrategias altruistas mantenidas en el repertorio de comportamientos de los animales mediante el mecanismo de selección por parentesco. De entre ellos destacaremos el altruismo que presentan determinados individuos de especies de insectos eusociales como las abejas, hormigas, avispas y termitas (O. HYMENOPTERA y O. ISOPTERA) (Wilson, 1971), que son estériles, el máximo altruismo posible, y cooperan en el mantenimiento de las larvas de la colonia. En el caso de HYMENOPTERA, esto es posible dada la peculiar composición genética del Orden, la haplodiploidía, y la peculiar proporción de sexos en la colonia, 3:1 a favor de las hembras. La haplodiploidía y el mecanismo de reproducción hace que las hembras de la especie tengan un grado de parentesco con sus hermanas del 0.75 mientras que con sus hermanos e hipotéticos hijos sólo tengan 0.25. Siendo así las relaciones de parentesco genético, lo más rentable en términos de beneficios genéticos para una hembra es invertir en el cuidado de

sus hermanas y permanecer estériles. Las especies del Orden ISOPTERA, a diferencia de los anteriores son diploides pero presentan una serie de propiedades que hacen posible mantener unas relaciones de parentesco genético que favorezca el altruismo reproductor. Estos mecanismos son una alternancia entre ciclos de endogamia y exogamia y también la acumulación de gran parte del genoma en los cromosomas sexuales. Ambos mecanismos se traducen finalmente en un efecto similar a la haplodiploidía en el sentido de que incrementan las relaciones de parentesco entre hermanos. Finalmente, los áfidos (O. HEMIPTERA - S.O. HOMOPTERA) (Aoki, 1983) presentan también individuos especializados con comportamientos altruistas en la defensa de la colonia frente a depredadores. La característica genética fundamental de este Orden es que las colonias de muchas de sus especies están formadas por individuos clónicos (grado de parentesco 1) (Itô, 1989; Foster, 1990).

Existen también ejemplos de comportamientos altruistas entre individuos genéticamente relacionados pero ajenos a los fenómenos de eusocialidad. En general, los costes del altruismo en estos casos son menores a los presentes en los comportamientos de altruismo reproductor, el suicidio o la defensa que se observan en los insectos eusociales. Una muestra de ellos es, por ejemplo, el cuidado parental[9], la cooperación entre hermanos del pavo americano para la reproducción, llamadas de alerta ante la presencia de depredadores en los perros de las praderas, comportamientos de crianza comunal en muchas especies de aves (Hidalgo, 1994).

Con lo anteriormente expuesto se justifica la estabilización de comportamientos altruistas entre individuos genéticamente relacionados. Pero es frecuente encontrar estos comportamientos también entre individuos no relacionados. Para la justificación de esta presencia de altruismo en grupos sociales de individuos no emparentados se proponen mecanismos que incrementan la eficacia inclusiva de los actores mediante el incremento de la eficacia Darwinista individual. Los mecanismos que se incluyen dentro de este grupo engloban el altruismo recíproco y el altruismo por retorno de beneficio (Trivers, 1985). La idea subyacente en estos mecanismos consiste en que el actor del acto altruista recibe una recompensa por dicho acto.

Así tenemos por ejemplo el mecanismo denominado de *beneficios indirectos o retorno de beneficios*. Un ejemplo estudiado en profundidad es la relación mutualista de limpieza que existe entre la familia Labridae (peces labroides) que limpian ectoparásitos de otras especies de peces coralinos. En esta relación, además de una cooperación honesta en la que los individuos limpiadores no engañan al hospedador ingiriendo tejido del mismo en vez

[9] El cuidado parental es un caso particular de comportamiento cooperativo altruista mostrado hacia individuos genéticamente relacionados. Consisten en la inversión que la generación parental hace en el cuidado de la generación filial. Aunque este comportamiento suele considerarse ajeno a los mecanismos de selección de parentesco, dado que afecta directamente a la eficacia darwinista de los actores, hemos optado por incluirlo en este apartado pues es consistente con la definición hecha de cooperación hacia individuos genéticamente relacionados.

de ectoparásitos y el hospedador se deja limpiar desatendiendo su propia búsqueda de alimento y no capturando a los limpiadores (Poulin y Vickery, 1995), existe otro comportamiento de señalización de presencia de depredadores considerado altruista. En él, el hospedador avisa a su limpiador de la presencia de depredadores perdiendo tiempo en la reacción de huida a cambio de que una vez pasado el peligro, el limpiador repita posteriormente el comportamiento de limpieza. También existen individuos con comportamientos altruistas en los casos de cría comunal que se encuentran en especies de aves. En ellas hay individuos que cooperan en la crianza de la prole ajena recibiendo a cambio beneficios de diversa índole, como son la adquisición de experiencia en el cuidado parental, beneficios por la pertenencia a un grupo, posibilidades de heredar parcelas de territorio, etc. (Reyer, 1984; 1986).

2.7 Altruismo recíproco.

El *altruismo recíproco* o *Reciprocidad* es otra forma de compensar el coste del altruismo encontrado entre individuos no relacionados genéticamente. Originalmente fue definido como el mecanismo que produce un incremento neto en éxito reproductor en un individuo por el hecho de ayudar a otro con un coste pequeño cuando este otro devuelve posteriormente la ayuda recibida (Trivers, 1971). En este sentido, el carácter altruista lo es sólo temporalmente y los costes del mismo son menores que la ganancia que se obtendrá después cuando se produzca la reciprocidad. Un análisis de este mecanismo descubre las condiciones que deben presentar las estrategias recíprocas para garantizar la estabilidad del comportamiento cooperativo altruista (Axelrod y Hamilton, 1981).

i) debe existir suficiente número de encuentros dos-a-dos para permitir, por un lado una correcta asignación de roles entre los miembros del grupo así como para asegurar que los actores del acto altruista reciban el beneficio de la reciprocación. Esta condición es conocida como *viscosidad.*

ii) el beneficio obtenido de recibir ayuda debe ser superior al coste de ayudar.

iii) los actores altruistas deben poseer un mecanismo de reconocimiento fiable para poder reconocer a los individuos que no se comportan de forma recíproca. Una vez reconocidos éstos, no volverán a ser ayudados. En las relaciones mutualistas y simbióticas, no suelen ser necesarios mecanismos específicos para el reconocimiento del engaño debido a que generalmente huésped y hospedador viven en contacto permanente. En otras situaciones, se han desarrollado mecanismos sensoriales de reconocimiento. Un caso extremo lo constituye el

sistema de reconocimiento de rostros presente en la especie humana. El cerebro humano ha dedicado una estructura para el reconocimiento de rostros[10].

Ejemplos de este mecanismo de reciprocidad son muy frecuentes en la naturaleza y se encuentran en una gran variedad de taxones. El comportamiento de regurgitación de sangre encontrado en una especie de vampiros tropicales (*Desmodus rotundus*) atiende a este mecanismo (Wilkinson, 1984). Estos murciélagos son animales coloniales que viven en grupos estables. Un aspecto crítico de su supervivencia es el hallazgo de fuentes de alimentación. Han desarrollado un mecanismo de reciprocidad mediante el cual un individuo que vuelve al nido tras haber encontrado alimento, frecuentemente regurgita parte de él hacia algún compañero del grupo. Los análisis de esta estrategia mostrada por individuos en libertad demuestran el cumplimiento de los requisitos antes mencionados. Concretamente, al formar grupos coloniales estables garantizan la condición (1), el efecto de una regurgitación sobre la supervivencia del donante es menor que la ganancia del receptor (2) y por último, un estudio experimental de captura y aislamiento de ejemplares en libertad demostró la eficacia del sistema de reconocimiento (Wilkinson, 1984). Otro ejemplo de reciprocidad interesante lo constituye el comportamiento de inspección de depredadores presente en los peces guppy (*Poecilia reticulata*) (Dugatkin, 1988; 1991) y en los peces espinosos (*Gasterosteus aculeatus*) (Milinski, 1987). Ante la presencia de depredadores, estos peces forman parejas de inspección de forma que se aproximan hacia el depredador para obtener mayor información acerca del mismo. Se postula que esta inspección confiere ventajas a la hora de escapar a los intentos de depredación (Pitcher, *et al.*, 1986).

Este mecanismo de reciprocidad es el que postulamos en esta tesis como estabilizador de estrategias altruistas en un colectivo de robots. No hay en la literatura del área ningún planteamiento que reúna las condiciones que se proponen en esta tesis, acerca de la solución al problema de la inestabilidad del altruismo. Entre los trabajos más próximos encontramos varias circunstancias que los separan de nuestro enfoque.

En primer lugar, los trabajos de (Balch y Arkin, 1994; Goss y Deneubourg, 1992; Kube y Zhang, 1993b; Steels, 1994a) desarrollan sistemas multiagente que carecen de capacidad de aprendizaje sobre los comportamientos cooperativos. En estos casos, la cooperación es una estrategia preprogramada, fijada en la arquitectura de comportamientos de los agentes con lo que no existen problemas de estabilidad. La ausencia de adaptabilidad, sin embargo, condiciona el ambiente en el que se pueden desenvolver estos sistemas. Los cambios en las condiciones de trabajo que repercutan en la adecuación de los comportamientos cooperativos se traducirán en detrimentos irreversibles del rendimiento global.

[10] De hecho, las personas con lesiones en estas zonas cerebrales sufren de prosopagnosia, enfermedad que se manifiesta con la incapacidad de recordar rostros.

Otro grupo de trabajos, implementa las estrategias cooperativas de forma que tanto el actor como el receptor de las mismas se beneficia inmediatamente. En este sentido son consideradas como mutualistas en lugar de altruistas, no existiendo entonces problemas de estabilidad (Tan, 1993; Mataric, 1995; Parker, 1994b; Weiss, 1995).

Finalmente, otra serie de trabajos desarrolla sistemas multiagente con evaluación global del rendimiento de los agentes (Murciano y Millán, 1996). La evaluación que proponen asemeja los mecanismos de selección por parentesco en el sentido de que los agentes individualmente asumen como propio el beneficio del colectivo. A diferencia de las evaluaciones locales que se proponen en esta tesis, un sistema de evaluación global provoca una cierta pérdida de autonomía de los agentes pues debe implementarse un sistema de supervisión que evalúe la acción conjunta del equipo de trabajo y comunique a todos los agentes los progresos en su rendimiento. Esta última necesidad incrementa los requerimientos de comunicación del sistema. En (Murciano, Millán y Zamora, 1997) se discute las ventajas e inconvenientes de las evaluaciones locales y globales del rendimiento aplicando dichos procedimientos en un modelo de especialización en sistemas multiagente.

Es necesario señalar que el problema planteado de la estabilización del altruismo en la naturaleza presenta cierta similitud, cuestionada por algunos autores (Stephens *et al.*, 1995), con un conocido modelo de la Teoría de Juegos del Dilema del Prisionero. En el Dilema del Prisionero (Axelrod, 1984; Axelrod y Hamilton, 1981;Rapoport y Chammah, 1965), dos jugadores se enfrentan teniendo que elegir entre dos posibles jugadas: cooperar (C) o defraudar (D). Los beneficios obtenidos por un jugador en cada jugada no sólo dependen de su elección sino que también dependen de la elección que haya hecho el contrincante. Así se establece la matriz de pago de la Tabla 2.2. De esta matriz viene el dilema. Si los jugadores se enfrentan un sola vez, la mejor estrategia posible, y a la par la única estable, es defraudar. Ahora bien, si ambos jugadores defraudan obtienen beneficios menores que si ambos hubieran cooperado. La situación cambia si el dilema del prisionero es iterado, es decir, si hay una probabilidad no nula (denominada ω), de que los mismos jugadores se enfrenten una vez más. En éste caso, una estrategia atractiva como solución del problema es la estrategia conocida como Tit For Tat (TFT) (Axelrod y Hamilton, 1981). Para una revisión del comportamiento de distintas estrategias en un torneo de simulación se puede consultar (Axelrod, 1984). Esta estrategia consiste en cooperar en la primera jugada y repetir lo que haya jugado el oponente en las siguientes jugadas. La estrategia TFT es considerada como una estrategia robusta (obtiene muy buenos resultados cuando se enfrenta a una gran variedad de estrategias) y además posee una serie de características que hacen de ella una estrategia muy atractiva:

i) Es una estrategia "bondadosa", es decir coopera en el primer movimiento del juego y nunca es la primera en no cooperar.

ii) Es una estrategia "provocable", es decir, responde con D cuando el oponente ha jugado D en la jugada anterior.

iii) Es una estrategia "no rencorosa", en el sentido de que si su oponente coopera después de haber defraudado, olvida anteriores defraudes y responde cooperando.

iv) es una estrategia fácil de entender por el resto de estrategias oponentes e implementable de forma muy sencilla.

Estas características hacen que TFT induzca a la cooperación a muchas otras estrategias, evite que se aprovechen de ella y sea capaz de restaurar la cooperación si el oponente, ocasionalmente o por error, ha defraudado.

Las características descritas de la estrategia TFT, la hacen especialmente atractiva para su implementación en la arquitectura de agentes autónomos. En el capítulo 4 presentaremos la arquitectura de reciprocidad (AREA), con una descripción de las condiciones y objetivos que condicionan las elecciones realizadas en el diseño de dicha arquitectura. En particular nos referiremos a las condiciones de estabilidad de la estrategia TFT.

Tabla 2.2. Tabla de pago del dilema del prisionero. El juego se define por las siguientes desigualdades: T>R>P>S y R>(S+T)/2. Por ejemplo, una combinación válida de valores numéricos para este dilema sería $R = 3$, $S = 0$, $T = 5$ y $P = 1$. (Axelrod y Hamilton, 1981)

		Jugador B	
		COOPERAR	DEFRAUDAR
Jugador A	COOPERAR	***R***	***S***
	DEFRAUDAR	***T***	***P***

2.8 Aprendizaje colectivo.

En el comportamiento animal es frecuente encontrar procesos de toma de decisiones que implican la selección de una alternativa de comportamiento. Un ejemplo, estudiado en profundidad, se refiere a la decisión que toman los animales sobre cambiar o no de lugar de forrajeo en función de la rentabilidad de cada una de las zonas disponibles. El rendimiento que obtiene de las elecciones realizadas depende en gran medida de las elecciones que realice el resto de individuos de su entorno. Así, si existen dos fuentes de alimentación con

distinta rentabilidad, y a su vez ésta depende del número de animales forrajeando en cada una de ellas, el problema es seleccionar aquella que reporta mayor beneficio, individual y colectivo. Como ya vimos en apartados anteriores, el criterio que dirige esta decisión es el criterio de optimalidad, es decir, el animal debe seleccionar la alternativa que confiera mayor rentabilidad, y debe hacerlo a partir del desconocimiento de las características del ambiente. Para ello deben contar con algún mecanismo de aprendizaje que, tras una evaluación de las distintas alternativas, posibilite la selección de la más rentable. Tras la aplicación de este mecanismo, los animales seleccionarán individualmente la alternativa de comportamiento que colectivamente alcance lo que se conoce como una *distribución libre ideal* (IFD) (Milinski y Parker, 1993).

La dependencia del rendimiento individual de un individuo de la actitud de las decisiones tomadas por otros, hace que el aprendizaje tenga un carácter colectivo. En este sentido, existe cierto paralelismo con la idea de la clasificación de los procesos adaptativos realizada en el apartado 2.3.4. Allí se argumenta que el proceso de aprendizaje es un proceso único, pero con dos vertientes según la escala de tiempo en la que opere. Así los animales usan mecanismos de adaptación fenotípica, como el aprendizaje, para seleccionar comportamientos en tiempo somático, bajo el criterio de maximización de la eficacia individual. La evolución se encarga de seleccionar comportamientos en tiempo evolutivo, evaluando una medida de eficacia inclusiva, que depende del éxito individual y también del éxito de los individuos genéticamente relacionados.

En las propuestas de modelos de aprendizaje biológicos para modificación del comportamiento animal, se deben tener en cuenta una serie de características fundamentales, comunes a toda regla de aprendizaje, individual o colectiva. Siguiendo a Maynard-Smith (1984), estas características se pueden enumerar como sigue:

1. Propiedad de la "Suma relativa de Beneficios" (RPS)

   Tras un periodo suficientemente largo de tiempo, que permita una estimación correcta de las rentabilidades relativas de cada alternativa posible de comportamiento, la probabilidad de mostrar cada una de ellas deberá ser igual al total de beneficio obtenido hasta el momento por la misma dividido por el total de beneficios obtenidos al mostrar cualquiera de las opciones posibles.

2. No extinción de alternativas

   Puesto que el ambiente puede cambiar, ninguna de las probabilidades de mostrar alguna de las alternativas de comportamiento debe anularse.

3. Probabilidades iniciales

   Los animales deben tener, a priori, unas probabilidades no nulas de mostrar las diferentes alternativas de comportamiento.

4. Factor de descuento.

   Los beneficios recibidos como consecuencia de las últimas acciones deben tener un efecto mayor sobre el comportamiento que los recibidos en tiempos anteriores.

A esta lista de características debemos añadir la necesidad de que exista un mecanismo que disminuya la exploración en función de varios factores; horizonte de tiempo total del individuo (explotación más exploración), variabilidad del entorno y rentabilidad acumulada. El proceso de exploración hace que el animal no se comporte óptimamente en el sentido más estricto, por lo que el mecanismo de aprendizaje deberá tenerlo en cuenta para poder alcanzar un balance óptimo entre exploración y explotación.

A continuación mostramos alguna de las reglas que se han propuesto en la literatura que cumplen con los criterios anteriores.

- *Regla RPS*:

La primera de ellas, es la propuesta por Harley (1981) denominada RPS. Esta regla, en el supuesto de dos posibles alternativas de comportamiento, A y B, se define como:

sean r_a y $r_b > 0$ los valores residuales, sea $0 < x < 1$ el factor de memoria y sea $b_i\,(t)$ el beneficio obtenido por mostrar el comportamiento $i \in \{A, B\}$ en el tiempo $t \in N$. Si en el tiempo cero, se inicializa el beneficio acumulado para las i alternativas $S_i(0) = r_i$ y para los $t > 0$ se actualiza esta estimación según

$$S_i\,(t) = x\,S\,(t-1) + (1-x)\,r_i + b_i\,(t),$$

la probabilidad de seleccionar en el tiempo t una alternativa $i \in \{A, B\}$ vendrá dada por la expresión

$$f_i\,(t) = \frac{S_i\;(t-1)}{S_A\;(t-1) + S_B\;(t-1)}$$

Es decir, de acuerdo a esta expresión, el animal elegirá más frecuentemente aquella alternativa de comportamiento que hasta el momento le haya proporcionado mayores beneficios, pero sólo en proporción al beneficio acumulado total.

- *Regla de muestreo por error.*

Otra regla de aprendizaje que contempla los mismos criterios, es la regla propuesta por Thuijsman *et al.* (1995) denominada *muestreo por error*. Esta se define, para un comportamiento con dos alternativas (A y B), como:

sea $\alpha \in (0,1)$ el factor de descuento, $e \in (0,1)$ factor de exploración, sea $a(t) \in \{A, B\}$ la alternativa seleccionada y $r(t) \in R$ el beneficio obtenido en el tiempo $t \in \{1,2,3,\ldots\}$. Se inicializa el nivel crítico $cl(0) = 0$ y para $t > 0$ se actualiza esta estimación mediante la expresión

$$cl(t+1) = \alpha\ cl(t) + (1-\alpha)\ r(t)\,.$$

Sea A_e la acción mixta, es decir, elige la alternativa A con probabilidad $(1-e)$ y B con la probabilidad complementaria (para B_e se usa una definición similar). La estrategia *error de muestreo* consiste en realizar la siguiente combinación temporal de elecciones:

para $t = 0$ elige $A_{0.5}$,

para $t = 1$ elige $a(0)_e$,

para $t \geq 2$ elige $a(t-1)_e$ si $a(t-1) \neq a(t-2)$ y $r(t-1) > cl(t-1)$.
$a(t-2)_e$ en caso contrario.

Tras realizar su primera elección al azar, el animal continuará mostrando dicha alternativa con una probabilidad $(1 - e)$ y explorará la contraria con probabilidad e. Si tras una exploración, recibe un beneficio menor que el esperado (que corresponde con el nivel crítico $cl(t)$, retorna a la alternativa anterior y permanece en ella de nuevo con una probabilidad $(1 - e)$. La actualización dinámica del nivel crítico cl, hace que el último beneficio obtenido tenga mayor importancia en la decisión que otros más alejados en el tiempo.

En el capítulo 4 describiremos la regla de aprendizaje implementada en la arquitectura de nuestros agentes autónomos que presenta una clara inspiración biológica. Es un mecanismo de aprendizaje colectivo que, al igual que los descritos anteriormente, utiliza una medida del beneficio de sus acciones para actualizar su criterio de selección de acciones. Guarda cierta relación con las reglas de Thuijsman *et al* (1995) y de Harley (1981). Durante su definición, en el capítulo 4, se expondrán las similitudes y diferencias. En el diseño del algoritmo se han tenido en cuenta los mismos principios que los descritos anteriormente. Esto es,

- la probabilidad de mostrar un determinado comportamiento es proporcional al beneficio obtenido por dicho comportamiento.
- se tiene un factor de descuento que pondera el efecto de anteriores rendimientos en la decisión actual primando los obtenidos más recientemente.
- el factor de exploración, más elaborado que el que presentado en las anteriores reglas varía de forma dinámica en función del grado de estabilización del rendimiento del agente y del rendimiento obtenido en las exploraciones.

El algoritmo de aprendizaje colectivo por refuerzo también cumple los requerimientos de los procesos adaptativos que se proponen en el apartado 2.3.1.

- El aprendizaje es incremental. A medida que el agente seleccione una alternativa de comportamiento, su rendimiento irá incrementándose. Además, el agente aprende a la vez que desarrolla la tarea, esto es, no se distingue una fase de aprendizaje y otra de explotación.
- El aprendizaje se concentra en el comportamiento relevante evitando aprender sobre otras facetas menos relevantes que pueden ser implementadas de forma reactiva.
- El proceso es no-supervisado. No se suministra al agente con información acerca de la dirección que debe seguir el proceso de aprendizaje.

2.9 Sumario del capítulo.

En este capítulo se presentan los fundamentos sobre los que se apoya la tesis. Se describe el concepto de agente autónomo discutiéndose sobre sus propiedades y sobre los objetivos perseguidos en el diseño de sus arquitecturas. Se plantea que la naturaleza y la robótica comparten objetivos como la optimalidad y la adaptabilidad, y comparten también vías para alcanzarlos, como es la inclusión de comportamientos sociales modificables mediante aprendizaje. Se describen estos conceptos, optimalidad, adaptabilidad y socialidad, y se presentan los mecanismos de adaptación en los seres vivos, en sus vertientes de aprendizaje y evolución, y la modelización de los mismos que se realiza en sistemas artificiales.

La conjunción en un mismo sistema, de mecanismos de aprendizaje con evaluación local, para conseguir la adaptación a distintos ambientes, y de comportamientos cooperativos altruistas, para incrementar el rendimiento del sistema, provoca la aparición de problemas de estabilidad de la cooperación. Se enuncia este problema de la estabilidad del altruismo y se enumeran los mecanismos presentes en la naturaleza para solucionarlo. Se propone el mecanismo denominado altruismo recíproco como solución, inspirada en la biología, para el mismo problema en el contexto de los sistemas multiagente.

Por último, el capítulo expone las propiedades de las reglas de aprendizaje biológicas en el contexto de aprendizaje en colectivos, ilustrándolas con varios ejemplos, discutiendo la relación que el aprendizaje colectivo de refuerzo tiene con las mismas.

Capítulo 3

Trabajos Relacionados

La presente tesis se relaciona con el diseño de sistemas multiagente cooperativos, compuestos por agentes autónomos que cuentan con capacidad de aprendizaje a partir de su propia experiencia. En este sentido, el trabajo que aquí se presenta se relaciona con la robótica colectiva en general, y en particular con aquellos sistemas que cuentan con capacidad de adaptación. En este capítulo se presentan los trabajos relacionados con los sistemas multiagente, comenzando con una breve revisión de los trabajos en los que la cooperación es una propiedad preprogramada que emerge de la interacción entre los agentes y no presentan capacidad de aprendizaje. Se revisan a continuación los sistemas multiagente con capacidad de aprendizaje y se señalan las diferencias que presentan con nuestro abordaje.

Una línea de investigación seguida frecuentemente en el contexto de los sistemas multiagente se fundamenta en la observación del comportamiento social de los animales, en especial, la observación de la compleja organización social de los insectos eusociales. En esta línea se encuadran los sistemas multiagente desarrollados por Deneubourgh *et al.* (1991), Deneubourgh, Theraulaz y Beckers (1992) y Goss y Deneubourg (1992). Se caracterizan por la ausencia de procesos de aprendizaje que posibiliten la modificación de los comportamientos en función de los requerimientos ambientales. Utilizan reglas de funcionamiento preprogramadas que ofrecen soluciones eficaces para las condiciones previstas en el diseño, pero su eficacia depende de la inmutabilidad de dichas condiciones y del conocimiento previo de las mismas. Los trabajos de Colorni, Dorigo, y Maniezzo

(1992) y Corbara, Drogoul, Fresneau y Ladande (1992), y Theraulaz, Goss, Gervet y Deneubourg (1991) también se fundamentan en la eusocialidad de los insectos careciendo, al igual que los anteriores de procesos de adaptación. La implementación física o aplicación a problemas reales de algunos modelos de inspiración eusocial puede encontrarse en Beckers, Holland y Deneubourg (1994) y Deneubourg, Clip y Camazine (1994).

Otros modelos cooperativos sin aprendizaje son los presentados por Drogoul y Ferber (1993) aplicados a tareas de recogida y almacenamiento de objetos. Estos modelos reciben el nombre genérico de Tom Thumb, los cuales, en sus versiones I y II, poseen capacidades de comunicación limitadas al tipo de comunicación estigmérgica. Un modelo más elaborado es el Tom Thumb III, que presenta comunicación inespecífica permitiendo a los robots la formación de cadenas de señalización para incrementar la eficacia en la ejecución de la tarea. Al no existir mecanismo de aprendizaje alguno, la idoneidad de su propuesta depende de las condiciones del ambiente de trabajo, con lo que cambios en las mismas limitan la eficacia del sistema.

La utilización de comportamientos reactivos en arquitecturas basadas en esquemas para la ejecución de tareas cooperativas se muestra en (Arkin, 1989). En trabajos subsiguientes se discute sobre las ventajas e inconvenientes de la comunicación en el desarrollo de tareas cooperativas (Arkin, 1992; Balch y Arkin, 1994). Kube y Zhang (1993a; 1993b) presentan un sistema multiagente, basado en arquitecturas denominadas subsumidas (Brooks, 1986), que aborda la tarea del desplazamiento cooperativo de objetos pesados. Posteriores refinamientos de su modelo básico pueden encontrarse en (Kube y Zhang, 1994). El trabajo colectivo desarrollado en este último caso, es dependiente de las condiciones de trabajo, como el peso de las piezas y el número de agentes.

Los anteriores trabajos comparten la carencia de procesos de adaptación por lo que no permiten la modificación del comportamiento de los agentes en función de los requerimientos del ambiente o de la tarea. Su funcionamiento se basa en el diseño de secuencias de comportamiento útiles para el funcionamiento de los agentes en ambientes conocidos y estáticos. La serie de trabajos que se exponen a continuación incluyen algún mecanismo de aprendizaje que los distingue de los anteriores.

Relacionados con el campo de la vida artificial, se encuentran una serie de trabajos que utilizan técnicas genéticas para la evolución de los sistemas de control de colectivos de agentes. Generalmente, estas técnicas utilizan estrategias evolutivas basadas en un gran número de agentes que evalúan su comportamiento en sucesivas generaciones. Werner y Dyer (1993) presentan un modelo en el que se combinan redes neuronales artificiales y algoritmos genéticos. De la evolución de este sistema, emergen secuencias de comportamiento básicas útiles para el funcionamiento de los agentes simulados. Un abordaje similar, pero enfocado a la evolución de estrategias de comunicación en colectivos de agentes se presenta en (MacLennan y Burghardt, 1994).

En los modelos de Numaoka y Takeuchi (1993) y Numaoka (1994) los agentes disponen de un repertorio de estrategias simples elegibles. Mediante métodos algorítmicos de instigación se obtienen combinaciones homogéneas de comportamientos en los agentes. Steels (1994a; 1994b) propone un modelo basado en la emergencia de comportamientos cooperativos dirigida por las necesidades impuestas por el ambiente. Una extensión de estos modelos consiste en la aplicación de un mecanismo selectivo ("selectron") entre nuevos comportamientos generados en modo aleatorio (Steels, 1994c). Dicho mecanismo faculta a los agentes para "sobrevivir" ante cambios en las condiciones del ambiente mediante su adaptación al mismo. Más recientemente (Steels, 1996) utiliza una representación de los comportamientos de los robots mediante acoplamientos entre parámetros (internos y externos) relevantes para la viabilidad del robot. La adecuación de las distintas funciones de acoplamiento es probada en un modelo de ecosistema en el que los agentes deben competir por los recursos energéticos.

Una interesante aplicación de modelos de robótica colectiva a la resolución del problema del viajante se puede encontrar en (Dorigo, Maniezzo y Colorni, 1995). Estos autores utilizan un sistema de comunicación basado en feromonas, similar al utilizado por las hormigas. Una serie de agentes recorre un grafo problema dejando el rastro de comunicación a su paso. De la iteración de este proceso resulta la optimización de la ruta solución al problema. Este sistema presenta grandes requerimientos de comunicación y el conocimiento global del problema para la generación de listas *tabú* para evitar repetidas visitas a los nodos del grafo. Dorigo y Gambardella (1995) introducen mejoras en el sistema anterior, implementando un algoritmo de aprendizaje por refuerzo del tipo Q-learning. La señal de "feromona" dejada por un agente en el recorrido es utilizada para realizar las estimaciones de utilidad de cada posible acción a partir de cada nodo. Los altos requerimientos de comunicación y la necesidad de reiterados recorridos por el grafo problema para el cómputo de señales de refuerzo globales, restringen su aplicabilidad en ambiente reales. Otra aplicación del algoritmo de aprendizaje por refuerzo se puede encontrar en el modelo de Tan (1993) donde utiliza un algoritmo Q-learning para el aprendizaje de secuencias de movimiento correctas. Los agentes cooperan intercambiando información acerca de mapeos correctos situación-acción con el objetivo de acelerar el proceso de aprendizaje. Es un caso particular de la relación entre el aprendizaje y la cooperación, en el sentido de que los agentes cooperan para aprender y no aprenden a cooperar.

En su tesis doctoral, Mataric (1993) realiza un interesante estudio acerca de la emergencia de comportamientos complejos a partir de la interacción de agentes con comportamientos básicos. Así, obtiene comportamientos colectivos como el agrupamiento y la dispersión a partir de comportamientos reactivos individuales. En trabajos posteriores presenta un algoritmo de aprendizaje heterogéneo que incluye refuerzo por múltiples objetivos (Mataric, 1994a; 1994b). El uso de estimadores de progreso en esos objetivos aumenta la velocidad de aprendizaje individual de los agentes. La inclusión de conceptos de

aprendizaje social en el algoritmo utilizado dotan al colectivo de sentido cooperativo explícito (Mataric, 1994c). Más recientemente (Mataric, 1996) presenta un algoritmo de aprendizaje utilizado en una tarea de recogida de objetos, en la que los agentes seleccionan individualmente acciones a partir de señales de refuerzo individual. Es interesante notar, que la utilización de señales de refuerzo individuales para aprender la estrategia colectiva óptima, requiere de la asunción de que la eficacia individual se corresponde con la eficacia colectiva, con lo que incrementos en el rendimiento individual de un agente resultan en un incremento del rendimiento colectivo. Esta asunción imposibilita el aprendizaje de estrategias sociales en las que los actos cooperativos sean altruistas, es decir, supongan un detrimento de la eficacia individual de los agentes.

Parker ha desarrollado una arquitectura (ALLIANCE) para el control de equipos heterogéneos de agentes. ALLIANCE contiene un conjunto de módulos motivacionales que son disparados por determinados vectores de entrada y que provocan el funcionamiento de bloques de comportamientos básicos. La activación de los módulos motivacionales depende de un cierto número de parámetros, que son prefijados por el diseñador. Este modelo ha sido probado en simulación de tareas de limpieza (Parker, 1994a). Una extensión de la arquitectura ALLIANCE con posibilidad de aprendizaje ha sido denominada L-ALLIANCE (Parker, 1994b). En ella los parámetros que controlan el disparo de los módulos motivacionales son modificables por aprendizaje. El tipo de aprendizaje escogido se basa en memoria adquirida en tandas de trabajo.

Murciano (1995) y Murciano y Millán (1996) presentan una variada colección de situaciones experimentales relacionadas con la solución cooperativa de una tarea de recogida de objetos. Sus agentes seleccionan distintas estrategias cooperativas, por medio de un algoritmo de refuerzo, obteniendo resultados óptimos en ambientes desconocidos y cambiantes. A partir de la evaluación global de la eficacia del colectivo, obtienen fenómenos de formación de cadenas de agentes señalizadores de la posición de los objetos para facilitar la captura de los mismos, y fenómenos en los que los agentes se especializan en tipos funcionales como señalizadores y recogedores, permitiendo la aparición de equilibrios intermedios dependientes del ambiente. En otra serie de trabajos Murciano (1995) y Murciano, Millán y Zamora (1997) presentan un colectivo de agentes con propiedades de especialización dependiente del ambiente. Los agentes se especializan para resolver eficientemente una tarea compleja de recogida y almacenamiento ordenado de objetos. Presentan un análisis comparado de las ventajas e inconvenientes de utilizar evaluaciones de la eficacia individual y colectiva.

Capítulo 4

Material y Métodos

El marco conceptual en el que se encuadra esta tesis plantea que la cooperación altruista en las sociedades animales comparte aspectos como su motivación y problemática con la cooperación en sistemas multiagente artificiales. El planteamiento experimental de la tesis debe, por lo tanto, reflejar este paralelismo entre ambos contextos, natural y artificial. Este capítulo describe el planteamiento experimental diseñado, que incluye descripciones del ambiente de trabajo y de la arquitectura de agente autónomo recíproco y por último, la descripción del algoritmo de aprendizaje utilizado. A lo largo de este proceso descriptivo se justificarán las elecciones realizadas con referencias, cuando proceda, a su fundamentación biológica.

4.1 Condiciones experimentales

Los experimentos realizados a lo largo de esta tesis, se han llevado a cabo mediante un sistema multiagente desarrollando una tarea colectiva. La tarea planteada, inspirada en la actividad de forrajeo de los animales, es la recogida de una serie de objetos distribuidos en el ambiente de trabajo. Ésta debe realizarse de acuerdo a los siguientes criterios:

1. Optimalidad

   Se persiguen rendimientos óptimos (o próximos al óptimo). En el caso que nos ocupa, el criterio de optimalidad se plasma en la cantidad de objetos recogidos por el sistema en un tiempo determinado. El mismo resultado se obtendría si cuantificaramos este criterio en términos de tiempo tardado en recoger un número n de objetos.

2. Adaptabilidad

   Se pretende que el sistema se adapte al ambiente en el que trabaja modificando su comportamiento con la intención de maximizar su rendimiento. Relacionado con este objetivo está el de la flexibilidad, esto es, una vez que el sistema se haya adaptado a un ambiente determinado, si éste cambia, el procedimiento de adaptación elegido debe permitir revertir de la actual configuración de comportamientos hacia aquella que consiga mejores rendimientos.

3. Altruismo

   Dado que trabajamos con un colectivo de agentes autónomos, el uso de estrategias cooperativas en su seno podrá incrementar la eficacia en el desarrollo de la tarea. Los agentes podrán comportarse de forma altruista, si las condiciones ambientales determinan que este comportamiento es ventajoso para el colectivo.

La observación de estos criterios se ve dificultada por el hecho de que los agentes están "situados" en el ambiente, percibiendo las características del mismo de forma parcial (no tienen un mapa global de la situación) y a partir de sus lecturas sensoriales (frecuentemente con ruido). Adicionalmente, el ambiente es modificado continuamente por la acción de los agentes, convirtiéndose por tanto en un ambiente extremadamente dinámico.

Este conjunto de criterios y de dificultades ha condicionado las elecciones realizadas en cuanto al diseño de la arquitectura de los agentes autónomos y del algoritmo de aprendizaje. Antes de entrar en esta descripción, presentaremos el ambiente experimental empleado.

El entorno de trabajo donde se desarrollarán los distintos experimentos de simulación ha sido diseñado mediante un simulador de agentes autónomos. Mediante este simulador es posible diseñar distintos ambientes de trabajo, diferenciándose entre sí en el número y posición de los obstáculos, en el número y propiedades de los agentes y en el número y distribución de los objetos que han de ser recolectados por el equipo de trabajo.

El mundo de trabajo es un espacio bidimensional de dimensiones 640 x 480 unidades (Figura 4.1). Al comienzo de cada ciclo de simulación, los agentes parten en busca de los objetos mientras navegan por el entorno evitando los obstáculos. Cuando un agente

localiza un objeto, se aproxima a él, lo recoge y vuelve a depositarlo en el almacén. Cada ciclo termina pasado un intervalo de tiempo $k_{t\_límite}$. En la Tabla 4.2 se muestran los parámetros usados en las simulaciones.

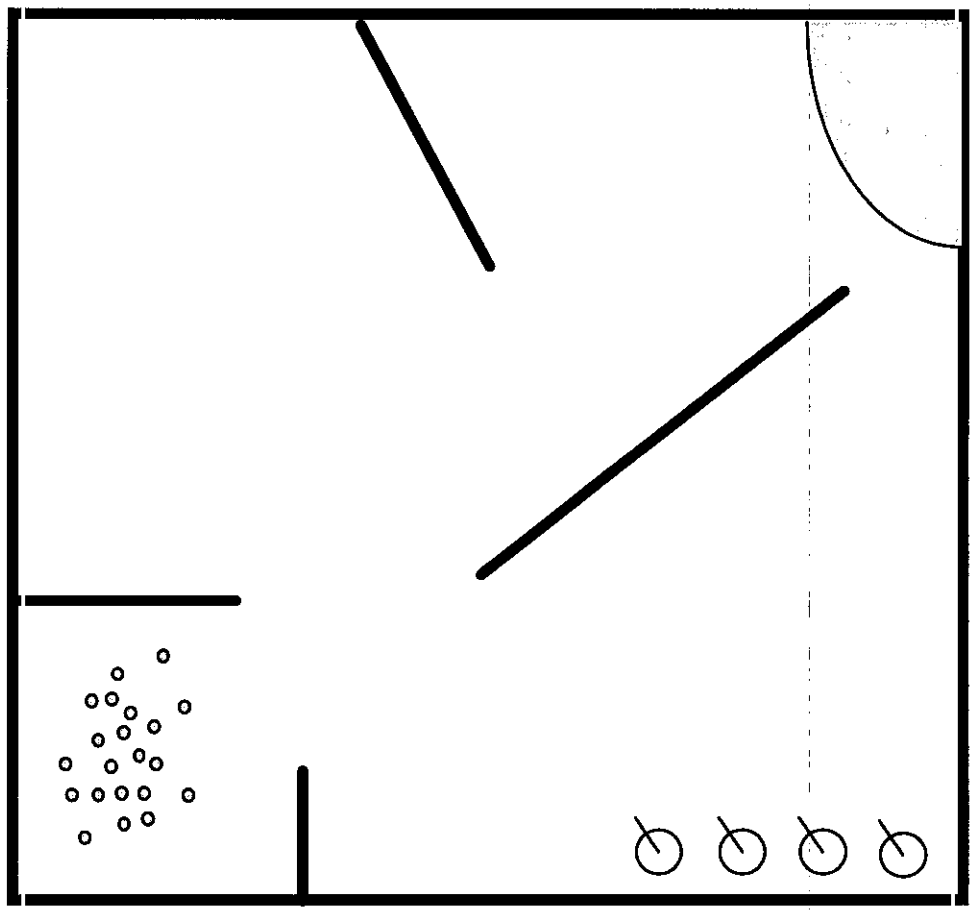


Figura 4.1. Ejemplo de mundo de trabajo en su posición inicial.

La fase experimental de la tesis se apoya fundamentalmente en la simulación, dadas las innumerables ventajas de velocidad y facilidad de implementación que se consiguen con esta metodología de trabajo. Se pueden generar ensayos de simulación con los que optimizar la búsqueda de parámetros adecuados para el control y aprendizaje de los agentes autónomos. Sin embargo, existe cierta controversia en la comunidad científica sobre las ventajas e inconvenientes del trabajo en agentes autónomos simulados frente al trabajo con implementaciones físicas de los mismos (Brooks, 1991a). Las críticas a la validez de los resultados de simulación se centran en la ausencia de interacciones "reales" con el mundo

físico. En su lugar se utilizan modelos del mundo que no recogen todas las características del ambiente sino que son reducciones de la realidad. Para incrementar la validez de los resultados de simulación, se han planteado dos estrategias de trabajo. Primeramente, en las simulaciones, se ha puesto especial énfasis en realizar una implementación realista del ambiente de trabajo, de la arquitectura de los agentes físicos y de las interacciones entorno-agente. Se persigue que las características de los agentes simulados se aproximen lo mejor posible a las propias de los robots reales. En especial se han tenido en cuenta las restricciones que el trabajo con robot físico impone en cuanto al rango y precisión de la lectura de sus sensores, errores en la comunicación e imprecisión de las acciones. Segundo, se ha probado la arquitectura de los agentes (sistemas sensoriales, efectores y de control) tanto en simulación como en la plataforma física del robot COOBOT, descrita en el capítulo 7. Se identifica así hasta qué punto el comportamiento de los agentes es equiparable en ambas plataformas.

4.2 Arquitectura de los agentes AREA

La arquitectura del agente autónomo AREA (acrónimo de Agente REcíproco Autónomo) es una arquitectura basada en comportamientos (Balch y Arkin, 1994; Brooks, 1986; Steels, 1994a). Esta arquitectura es similar a la realizada en los trabajos de Murciano (1995), Murciano y Millán (1996) y Murciano, Millán y Zamora (1997) en los aspectos del diseño de los comportamientos reactivos así como de los dispositivos sensoriales y motores. Sin embargo se diferencia en la inclusión de los comportamientos necesarios para la implementación de estrategias de cooperación recíproca.

En los sistemas basados en comportamiento, el conjunto de las posibles acciones se agrupan en unidades de más alto nivel denominadas comportamientos. Cada uno de ellos se dedica a una determinada competencia de la tarea global del agente. Las competencias que debe resolver el agente pueden ser de propósito general, como la evitación de obstáculos, o específicas de la tarea encomendada, en nuestro caso, coger un objeto. El diseño completo de la arquitectura debe contemplar los sistemas sensoriales que proporcionarán las percepciones, los sistemas efectores que ejecutarán las acciones, y el sistema de control que conectará percepciones y acciones y determinará la secuencia espacio-temporal de ejecución de los comportamientos.

4.2.1 Dispositivos sensoriales.

Los agentes cuentan con una serie de sensores que recabarán la información relevante para su funcionamiento. Se distinguen dos apartados según su relación con distintos sistemas funcionales del agente.

1. Sistema sensorial para la navegación.

   Este sistema sensorial extrae información próxima al agente para utilizarla en tareas de navegación por el mundo de trabajo. Cuenta para ello con dos tipos de sensores situados en distintas posiciones.

   1.1. Sensores de distancia que cuantifican la cercanía del agente hacia los obstáculos (un agente considera al resto de agentes como obstáculos). Son siete sensores situados en la parte frontal del agente en las posiciones -90°, -60°, -30°, 0°, +30, +60°, +90° respecto al eje central del robot. La activación de estas señales es directamente proporcional a la distancia a los obstáculos.

   1.2. Sensores de contacto para detectar colisiones entre el agente y otras partes del mundo (obstáculos y otros agentes). Están situados en las mismas posiciones que los anteriores.

2. Sensores para la localización de los objetos.

   Los agentes cuentan con un sensor de objetos localizado en la parte inferior de su estructura de forma que cualquier objeto situado a una distancia menor k_{obj} y dentro de un rango angular de [+22,5°, -22,5°] desde el eje del agente, activa este sensor. Una vez que el agente se aproxima a dicho objeto hasta una distancia inferior k_{rec} se activa un sensor de contacto indicando que el objeto está dentro del área de recogida.

4.2.2 Dispositivos Efectores.

El agente AREA cuenta con un sistema motor basado en dos ejes motrices los cuales giran independientemente. Los motores que controlan ambos ejes pueden estar en tres estados: inactivo o STOP, giro hacia adelante o AVA y giro hacia atrás o ATR. Las combinaciones de estos tres estados producen los 9 movimientos posibles que se muestran en la Tabla 4.1. Cada uno de esos movimientos desplaza el eje correspondiente una distancia fija k_m.

Finalmente los agentes cuentan con un sistema efector dedicado a la captura de los objetos.

Tabla 4.1. Combinaciones de movimientos del agente como combinación de los tres estados en que puede encontrarse cada uno de los motores que controlan cada eje motriz.

Eje derecho	Eje izquierdo	Movimiento
AVA	AVA	Adelante
AVA	ATR	Giro hacia la Izquierda total (GIT)
AVA	STOP	Giro a la Izquierda parcial (GIP)
ATR	AVA	Giro a la Derecha total (GDT)
ATR	ATR	Retroceso
ATR	STOP	Retroceso Derecha parcial (RPD)
STOP	AVA	Giro a la Derecha parcial (GDP)
STOP	ATR	Retroceso Izquierda parcial (RPI)
STOP	STOP	Parada

4.2.3 Sistema de comunicación.

Como hemos mencionado anteriormente, el trabajo del sistema multiagente deberá realizarse de forma cooperativa. La cooperación, en este caso, está basada en la comunicación explícita. Se deben por tanto incluir en la arquitectura de los agentes los dispositivos emisores y receptores necesarios para permitir la comunicación.

El sistema de comunicación utilizado por los agentes está basado en señales de infrarrojos (IR). Los agentes disponen de un emisor omnidireccional de infrarrojos y de 8 receptores de infrarrojos situados en una corona circular. Mediante ella, detectan las señales enviadas por otros agentes o por el almacén. La percepción de una señal de IR conlleva el conocimiento de la dirección y el contenido del mensaje que recibe. El sistema de comunicación está basado en una codificación binaria de los mensajes de forma que en un byte de información se incluye la identificación del agente emisor del mensaje (5 bits de mayor peso), y el tipo de mensaje (3 bits de menor peso). Hay tres tipos de mensajes que se intercambian los agentes. Estos son:

- Mensaje de *petición*: Los agentes en busca de objetos envían este tipo de mensajes para solicitar información de presencia de objetos a otros agentes.

- Mensaje de *donación*: Es la respuesta a una *petición*. Si un agente percibe directamente un objeto, responde a las peticiones informando que está percibiendo objetos.
- Mensaje de *relevo*: Cuando un agente comienza a percibir un objeto directamente y se dispone a pararse a señalizar, envía este mensaje. De ésta forma se evita que dos o más agentes están parados señalizando el mismo objeto o grupo de objetos.

El almacén emite continuamente una señal característica que sirve a los agentes para aproximarse a él a depositar los objetos que transportan.

La comunicación es explícita, de bajo coste y no dirigida, es decir los agentes no emiten los mensajes hacia ningún agente en concreto. Por ésta razón, cualquier agente dentro de un rango de distancia al emisor k_{com} percibirá cualquier señal emitida por éste. Dadas las características de la señal utilizada para la comunicación, los obstáculos impiden la propagación de las mismas, creando zonas de sombra entre los obstáculos, y el alcance de la comunicación es limitado en el espacio.

4.2.4 Sistema de control.

El sistema de control del agente debe realizar la conexión entre la parte sensorial y la parte motora de forma efectiva. Es el apartado más crítico de la arquitectura del agente. Se debe decidir qué información (situación) es relevante para qué acción o acciones y configurar así el conjunto de comportamientos. La existencia de distintos comportamientos, algunos de ellos posiblemente activados de forma simultánea, impone la necesidad de arbitrar algún mecanismo jerárquico, de prioridades o de competición entre los distintos comportamientos, que coordine temporalmente la ejecución de los mismos.

Al igual que en el comportamiento animal, se pueden distinguir dos tipos fundamentales de comportamientos en la arquitectura AREA.

1. Comportamientos reactivos:

   Este grupo de comportamientos, que en los animales recibe el nombre de *innatos*, es una serie de esquemas reactivos fijos que capacitan al agente para realizar las tareas de navegación y captura de objetos de forma individualizada. La eficacia de los estos comportamientos es independiente del ambiente en que se desenvuelvan los agentes. Dada la universalidad de las consecuencias de estas acciones, es posible realizar su preprogramación y proporcionarán a los agentes rendimientos aceptables desde el momento de su "nacimiento" en cualquier ambiente.

2. Comportamientos seleccionables/modificables:

   Este segundo grupo de comportamientos contiene aquellos que pueden modificarse a través de la experiencia. La adecuación de los mismos es dependiente del ambiente, que es desconocido, por lo que es imposible determinar a priori cual es la mejor configuración posible de los mismos. Por ello, estos comportamientos están sujetos a aprendizaje. En el sistema que se presenta, el comportamiento cooperativo es el que está sujeto a aprendizaje.

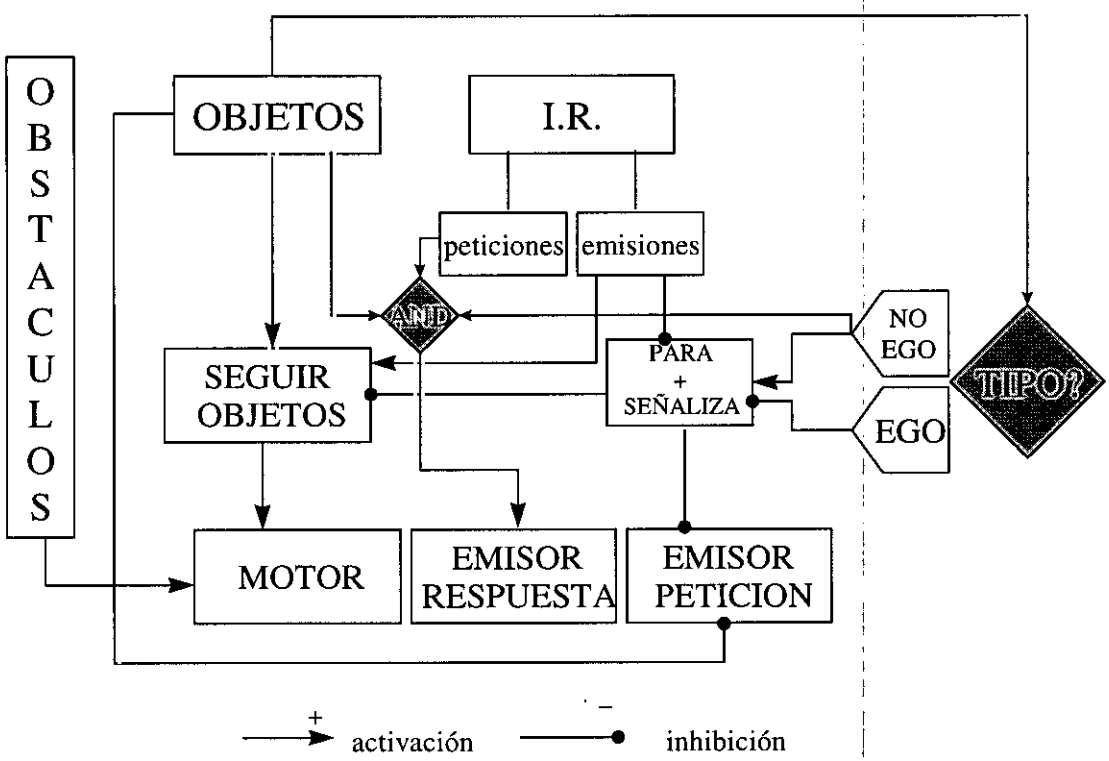


Figura 4.2. Gráfico resumen de los módulos de comportamiento del agente AREA. Por simplicidad no se muestra el diagrama de flujo del comportamiento recíproco TFT y se presenta una reducción del comportamiento cooperativo a las alternativas cooperar o no cooperar. Ver en el texto los detalles.

El control del agente es un proceso de tres etapas en las que el agente adquiere la información sensorial, desencadena un proceso interno de activación de comportamientos relacionados con la situación percibida y finalmente se ejecuta la acción. La información sensorial adquirida y el estado "motivacional" interno del agente, generalmente activan (o inhiben) de forma simultánea múltiples comportamientos (Figura 4.2). El nivel de

activación de todos ellos es sumado y normalizado produciéndose así la siguiente acción del agente. En este sentido, esta arquitectura es más próxima a la propuesta por Balch y Arkin (1994), Steels (1994a) o Murciano y Millán (1996) que a las arquitecturas conocidas como subsumidas (Brooks, 1986; Mataric, 1995; Parker, 1994).

Para el control de las decisiones de movimiento, se cuenta con tres unidades que representan la tendencia que cada motor tiene a estar en un estado determinado (giro hacia adelante, giro hacia atrás o parado). Estas unidades reciben conexiones excitatorias e inhibitorias desde los distintos módulos de comportamiento y desde los distintos sensores de distancia. Un ciclo de competición del tipo *winner-take-all* entre dichas unidades decide cual será la siguiente acción motora.

A continuación se ofrece la lista y descripción de los comportamientos reactivos fijos.

- **`Evitación-de-obstáculos:`** este comportamiento hace que el agente evite colisionar con los obstáculos. Existe un patrón de conexiones sensomotoras invertido, de forma que la percepción de obstáculos por uno de los laterales conecta inhibitoriamente con el lateral opuesto y excitatoriamente con el lateral correspondiente. La eficacia de estas conexiones laterales y un umbral de disparo adecuado hace que el movimiento del agente sea de trazado suave.

- **`Exploración:`** este comportamiento complementa al anterior de forma que cuando el agente esté navegando por una zona libre de obstáculos, decide un movimiento aleatorio con una cierta probabilidad. Con este movimiento se consigue aumentar la zona de exploración pues se evita el que los agentes se muevan indefinidamente por rutas similares.

- **`Seguimiento-de-objetos:`** si un agente percibe un objeto, se aproxima hacia él. Este comportamiento está inhibido si el agente está transportando un objeto.

- **`Pérdida-de-objetos:`** si un agente percibe un objeto y repentinamente deja de percibirlo (bien porque el objeto se salga del rango de percepción de su sensor de objetos, o bien porque otro agente se haya interpuesto entre él y el objeto), el agente desencadena este comportamiento que consiste en una secuencia de giros totales a ambos lados hasta completar un arco de 120° (aproximadamente) o hasta que se recupere la percepción del objeto.

- **`Captura-de-objetos:`** si un agente, mediante el comportamiento de **`seguimiento-de-objetos,`** se aproxima a un objeto y entra en contacto con él, es decir, se activa su sensor de contacto de objetos, dispara este comportamiento que se compone de una parada de los motores de movimiento y entrada en funcionamiento del dispositivo motor de captura de objetos.

- **`Regreso-almacén:`** si un agente transporta un objeto, se dirige hacia al almacén mediante la señal de IR emitida por el almacén. Cuando el agente entra en la zona de almacén deposita el objeto.

Los comportamientos descritos son comportamientos no sociales. Mediante este repertorio, los agentes pueden desarrollar la tarea de recogida de objetos de forma individual, no colectivamente. Existen sin embargo otros dos comportamientos reactivos no modificables que tienen que ver con el comportamiento social. Ambos se relacionan con el control de los dispositivos de comunicación. Son los siguientes:

- **`Seguimiento-de-señales:`** si un agente no está transportando ningún objeto y recibe mensajes de donación, dispara un proceso de seguimiento de la señal hasta que comience a percibir directamente los objetos o deje de percibir esta señal.
- **`Petición-de-ayuda:`** cuando un agente está buscando objetos y no los percibe directamente, emite un mensaje de petición.

Además del conjunto de comportamientos anterior, los agentes cuentan con otro grupo de comportamientos aprendibles, que se encarga del control del comportamiento cooperativo de los agentes. Mediante éste nuevo grupo se consigue que el colectivo de agentes trabaje de forma cooperativa incrementándose así la eficacia del sistema.

- **`Donación-de-ayuda:`** si un agente percibe un objeto, se detiene y cuando recibe un mensaje de petición de ayuda, responde a la petición con este comportamiento emitiendo mensajes de donación. Este comportamiento continúa hasta que el agente recibe un mensaje de relevo o hasta que se alcanza un tiempo límite t_{lim}. Cada vez que el agente emite una donación de ayuda, reinicia el contador de tiempo límite de espera. [11]
- **`Relevo-de-señalizador:`** si un agente comienza a percibir un objeto, dispara este comportamiento mediante el cual emite durante un periodo corto de tiempo mensajes de relevo.

La ventaja de mostrar estos comportamientos depende del ambiente donde se desarrolle la tarea. Si el ambiente contiene objetos agrupados en zonas poco accesibles (Figura 4.3), la

[11] Si un agente recíproco (ver más adelante) agota su tiempo límite de señalización, inhibe su comportamiento `donación-de-ayuda` durante ese ciclo de recogida. Este agente desinhibe este comportamiento si recibe alguna señal de ayuda de otro agente.

mejor estrategia para el colectivo es que los agentes muestren los comportamientos anteriores de cooperación. Ahora bien, si el ambiente presenta los objetos uniformemente distribuidos y muy accesibles, estos comportamientos provocarán pérdidas de rendimiento. Como los agentes no conocen la distribución de objetos y desconocen su grado de accesibilidad, deben aprender cual es la mejor estrategia para el ambiente al que se enfrentan.

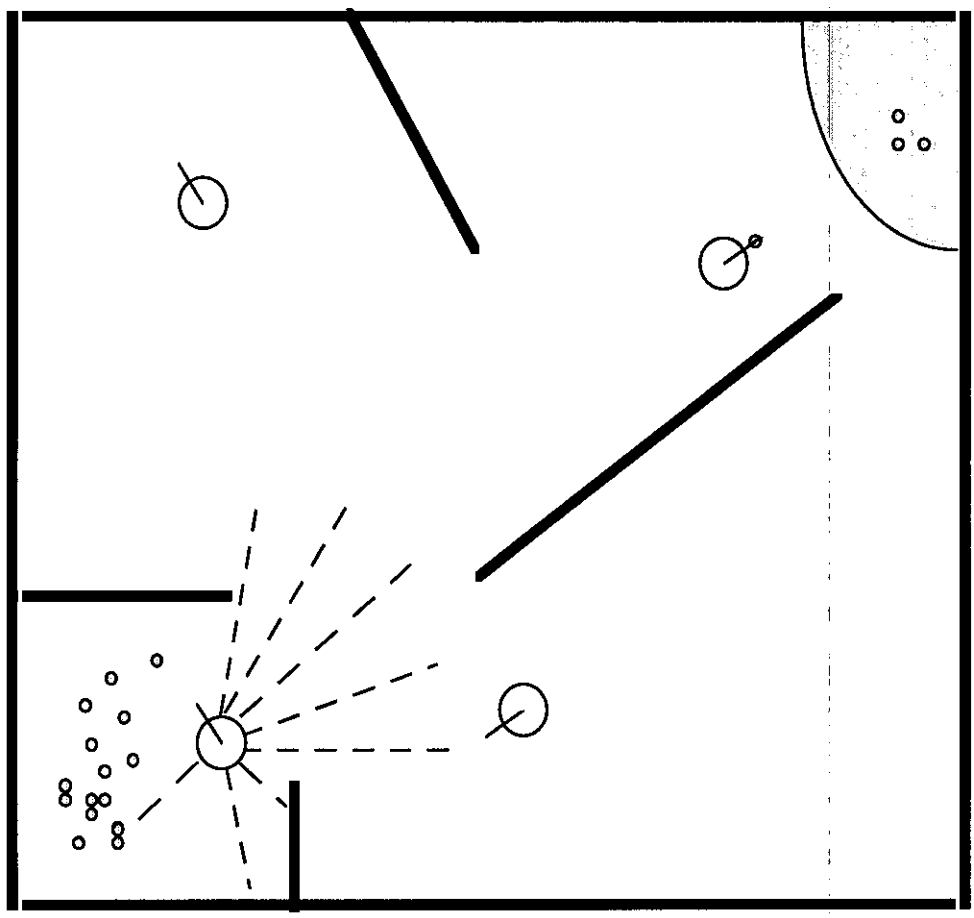


Figura 4.3. Ejemplo de la tarea de recogida en un ambiente con objetos agrupados

El aprendizaje se realiza sobre la tendencia que tienen los agentes a mostrar estos comportamientos. Esta tendencia está representada en una serie de variables $s_j(t)$, donde j representa las alternativas del comportamiento, que en el caso general es un comportamiento con tres alternativas: cooperar siempre, cooperar de modo recíproco o no

cooperar. La probabilidad de mostrar cada una de las alternativas del comportamiento cooperativo es proporcional al valor de las variables $s_j(t)$ (ver sección 4.3).

Si un agente decide mostrar estos comportamientos, diremos que se comporta de forma *altruista*. Si, por el contrario, decide no mostrarlos, diremos que ese agente es *egoísta*. La decisión de comportarse de una u otra forma se realiza de acuerdo a la rentabilidad de cada una de las opciones. Si los agentes tienen a maximizar el número de objetos que recoge (medida de refuerzo local), la estrategia más rentable es no cooperar, esto es, no perder tiempo en señalizar y, si las hubiera, usar las señales de objetos de otros agentes. Como hemos visto, esto se traduce en la desaparición del comportamiento altruista. Para poder mantener la cooperación en el sistema y conseguir así rendimientos óptimos, se debe implementar alguna de las posibles soluciones que se planteaban en el capítulo 2.

Nuestra propuesta de estabilización consiste en modificar la arquitectura de los agentes, para incluir los mecanismos necesarios para la implementación de la estrategia de *reciprocidad*, y conseguir así la estabilización del altruismo con el consiguiente incremento en el rendimiento del sistema. Así pues se incluye una tercera opción en el comportamiento cooperativo de los agentes, denominada reciprocidad. Un agente recíproco muestra los dos comportamientos anteriores (**`donación-de-ayuda`** y **`relevo-de-señalizador`**) pero de forma selectiva, esto es, sólo hacia los agentes que son también cooperativos evitando hacerlo hacia los agentes no cooperativos.

A modo de resumen, según el comportamiento cooperativo que seleccione un agente, podremos tener agentes de tres tipos.

1. *Agente egoísta* (no cooperativo).

   No muestra los comportamientos cooperativos **`donación-de-ayuda`** ni **`relevo-de-señalizador`.**

2. *Agente Altruista* (cooperativo).

   Muestra los comportamientos cooperativos **`donación-de-ayuda`** y **`relevo-de-señalizador`** incondicionalmente, esto es, hacia todos los agentes independientemente de si son o no cooperativos.

3. *Agente Recíproco* (cooperativo).

   Muestra los comportamientos cooperativos **`donación-de-ayuda`** y **`relevo-de-señalizador`**, sólo hacia otros agentes que sean también cooperativos. Para realizar la clasificación entre agentes según cooperen o no, los agentes recíprocos muestran la estrategia TFT, esto es, coopera en la primera interacción con un agente, independientemente de su tipo, y continuará haciéndolo siempre que este agente se catalogue, tras ese primer encuentro, como agente cooperador (ya sea altruista o recíproco). Mediante un comportamiento de **`reconocimiento-de-`**

señalizadores (ver más abajo) estos agentes determinan si el agente ayudado es cooperador (altruista o recíproco) o no cooperador (egoísta).

La estabilidad de esta estrategia se fundamenta en tres requisitos fundamentales:

i) Reconocimiento del engaño,

ii) suficiente número de interacciones entre los agentes,

iii) la ventaja de ser ayudado debe ser mayor que el coste de ayudar.

Un sistema que cumpla con estos tres requisitos será capaz de estabilizar la cooperación altruista en el seno del grupo.

Para el primero de ellos, el reconocimiento del engaño, proponemos la inclusión del siguiente comportamiento en la arquitectura de los agentes:

- **Reconocimiento-de-señalizadores:** este comportamiento permite a los agentes clasificar al resto de acuerdo a su comportamiento cooperativo. Los agentes muestran el comportamiento de **donación-de-ayuda** la primera vez que interaccionan con un agente. Si tras un periodo de tiempo t_{tft} no reciben ningún mensaje de relevo del agente que fue ayudado, éste es catalogado como agente no cooperativo (en otros términos, egoísta). Esta catalogación no es irreversible pues, en cualquier momento del ciclo de simulación, si un agente recibe, bien un mensaje de donación o bien un mensaje de relevo, cataloga (o recataloga en su caso) al emisor como agente cooperativo.

Este nuevo comportamiento permite a los agentes clasificar al resto como cooperativos o egoístas. Esta clasificación les permite mostrar la mencionada estrategia TFT. Esto es, cooperar en el primer encuentro con un agente y en los siguientes, repetir el comportamiento que tenga el contrincante.

La condición de *viscosidad* (ii) se cumplirá seleccionando adecuadamente el tiempo que transcurre entre iteraciones del aprendizaje, esto es, la duración de un ciclo de recogida determinado por la constante k_{t_limite}. Tiempos excesivamente cortos no permitirán una correcta asignación de roles mediante el comportamiento **reconocimiento-de-señalizadores.** Tampoco se podrá obtener la compensación del gasto de la cooperación pues no habrá tiempo suficiente para reciprocar, y por último, dado que los agentes egoístas se aprovechan del primer encuentro con cada uno de los recíprocos, una baja viscosidad puede imposibilitar el que los agentes recíprocos remonten esta desventaja inicial por medio de la cooperación recíproca. En el otro extremo, duraciones de ciclo de

recogida excesivamente largos afectarán a la velocidad de convergencia del algoritmo de aprendizaje. En el capítulo 6 analizaremos la influencia de la viscosidad sobre la estabilidad de la reciprocidad presentando una forma de estimar tiempos idóneos.

Por último, la tercera condición depende del ambiente y de la presencia, en la arquitectura de los agentes, de comportamientos adecuados para explotar la ventaja de la cooperación si ésta existe. Como hemos planteado, en ambientes donde los objetos estén uniformemente distribuidos, la cooperación no es ventajosa. En términos de estabilidad de la reciprocidad, diremos que no se cumple la condición iii) pues la ventaja de ser ayudado (localizar un objeto mediante un mensaje de donación) será menor que el coste de ayudar (tiempo parado señalizando). Si el ambiente presenta los objetos agrupados y poco accesibles, la cooperación es ventajosa pues el tiempo invertido señalizando es compensado con crece con la disminución del tiempo de búsqueda de objetos. La arquitectura de los agentes incluye los comportamientos necesarios para explotar esta ventaja de la cooperación. La presencia de los comportamientos reactivos sociales (**`seguimiento-de-señales`** y **`petición-de-ayuda`**) en el repertorio AREA posibilita la explotación de la ventaja de la cooperación.

En el capítulo 2 vimos las propiedades atractivas que tiene la estrategia de reciprocidad en el contexto de la cooperación animal. En el capítulo 5 veremos que un sistema multiagente con esta estrategia, es capaz de estabilizar la cooperación altruista cuando se cumplen las tres condiciones: el altruismo es ventajoso, la viscosidad del sistema es suficientemente alta y el mecanismo de reconocimiento del engaño es suficientemente fiable.

4.3 Aprendizaje

Un sistema de control puramente reactivo puede ser efectivo en ambientes conocidos y estáticos. Pero hay situaciones en las que el ambiente es desconocido o es variable. Además, una misma acción puede tener consecuencias radicalmente opuestas dependiendo del ambiente donde se produzcan. Por otro lado, la percepción que el agente tiene del mundo y las consecuencias de sus acciones sobre el mismo tienen imprecisiones que dificultan el trabajo con estos sistemas reactivos puros. Para solucionar estos problemas, es importante dotar a la arquitectura de los agentes, de mecanismos de aprendizaje que posibiliten la modificación del comportamiento a partir de la experiencia.

El mecanismo de aprendizaje se centra en el aspecto más relevante del funcionamiento de los agentes, evitando invertir en el aprendizaje de comportamientos que pueden ser implementados de forma reactiva. En consecuencia, el aprendizaje se realiza sobre el comportamiento cooperativo. Se ha implementado un mecanismo de adaptación al

ambiente, concretamente un mecanismo de aprendizaje por refuerzo, para que los agentes seleccionen la alternativa de comportamiento cooperativo (altruista, recíproca o egoísta) más rentable, cualquiera que sea el ambiente donde se desarrolle la tarea.

En el aprendizaje por refuerzo, el agente aprende de su propia experiencia. Mediante un mecanismo de *ensayo y error,* prueba distintas alternativas, evaluando la adecuación de las mismas para la situación en la que se encuentra. En su formulación básica (Kaelbling, 1990), presentada en el capítulo 2, los algoritmos de aprendizaje por refuerzo tratan de realizar un mapeo entre situaciones y acciones que permita, mediante un mecanismo de selección de acciones adecuado, maximizar una señal de refuerzo a lo largo del tiempo. En el caso que nos ocupa, es preciso reformular este planteamiento básico. Concretamente, el agente debe decidir entre mostrar el comportamiento cooperativo (enviar mensajes de donación y de relevo), hacerlo de forma condicional o no hacerlo, ante la situación de "percepción de un objeto". Por lo tanto, las acciones del agente se agrupan en comportamientos, y no es necesario realizar una definición explícita de las distintas situaciones en las que se puede encontrar un agente, puesto que la única situación relevante es la percepción de un objeto. Desde el punto de vista del agente (individual), el objetivo del aprendizaje es seleccionar la alternativa del comportamiento cooperativo que maximice el número de objetos recogidos por él, en un tiempo determinado. Desde el punto de vista del sistema completo (colectivo), el objetivo es alcanzar y estabilizar una distribución de comportamientos cooperativos en el equipo de trabajo que proporcione rendimientos cercanos al óptimo. En determinados ambientes, se tiene que ambos puntos de vista no coinciden. En el caso de los comportamientos egoístas, éstos son más rentables desde una perspectiva individual pero están en contraposición con el criterio de rentabilidad desde una perspectiva colectiva. La misma inconsistencia se tiene entre la percepción que un individuo altruista tiene de su bajo rendimiento individual, y la visión del colectivo que se ve beneficiado por la presencia de estos individuos.

El aprendizaje por refuerzo sufre de una serie de problemas que se engloban en lo que se conoce como *problemas de asignación de crédito.* Estos problemas reflejan la dificultad que un agente tiene de poder asignar correctamente a qué acciones corresponde las señales de refuerzo que se obtienen. En el caso que nos ocupa, este problema se ve influenciado por el carácter colectivo del aprendizaje. En este sentido, el rendimiento del comportamiento de un agente depende del comportamiento que muestre el resto del colectivo[12], cuestión que el agente desconoce completamente. Por ejemplo, el rendimiento de un agente con comportamiento altruista depende de si existe algún otro altruista en la población, y de su número. Si todos los agentes son altruistas, el rendimiento individual de cada uno de ellos es cercano al óptimo. A medida que disminuyen en número, el tiempo que un agente está parado señalizando se incrementa (tarda más tiempo en ser relevado) con el consiguiente detrimento de su rendimiento. Esta dependencia de la rentabilidad de

[12] En este sentido, el problema tratado en esta tesis puede ser considerado como un *juego dependiente de la frecuencia.*

los comportamientos hace que el agente perciba distintas señales de refuerzo al mostrar un mismo comportamiento encontrándose con el mencionado problema de asignación de créditos.

Por otro lado, el carácter probabilístico de los rendimientos obtenidos por los agentes incide también en la asignación correcta de créditos. En este sentido, se tiene con mucha frecuencia que la señal de refuerzo que un agente percibe en dos ciclos distintos, con la misma combinación de comportamientos en el colectivo, pueden ser diametralmente opuestos. La señal de refuerzo contiene una importante componente aleatoria que es dependiente de factores tan variados como el tiempo que tarda el colectivo en encontrar el primer objeto, el comportamiento que presenta el agente que llega en primer lugar a ese objeto, el número de agentes en el equipo, los errores en la comunicación, etc.

Finalmente, el mecanismo de aprendizaje debe superar una dificultad adicional, relacionada con la viabilidad inicial de los comportamientos cooperativos altruistas. En una población en la que todos los agentes muestran la alternativa de comportamiento egoísta, la instauración de la cooperación, aunque sea ventajosa, es difícil. Para que un agente aprecie la rentabilidad de esta estrategia, es necesario que aparezca en la población acompañado de al menos otro agente cooperativo. De no ser así, su intento de cooperar será aprovechado por el resto de agentes sin recibir ninguna compensación por el coste en el que incurre. En la naturaleza, se ha propuesto la participación de mecanismos de selección de parentesco en la fase inicial de instauración del altruismo, para garantizar la compensación del coste altruista mediante el beneficio de los individuos genéticamente relacionados. Otra forma de facilitar esta viabilidad inicial, es el que propone que los individuos altruistas aparecen apiñados, de forma que la mayor parte de las interacciones las realizan entre ellos. Ambos mecanismos están muy relacionados y puede que participen conjuntamente (Axelrod y Hamilton, 1981).

A pesar de las dificultades señaladas, el algoritmo de aprendizaje consigue estabilizar estrategias de comportamiento óptimas de forma dependiente del ambiente. En los capítulos 5 y 6 se muestran los resultados experimentales de la aplicación de este algoritmo.

4.3.1 Algoritmo de aprendizaje utilizado.

Definiciones previas:

En todas las definiciones se tiene que $i \in \{1,2,3,\ldots,n\}$ siendo n el número de agentes en el sistema y $1 \leq t \leq t_{lim}$.

1. Rendimiento individual

Se define el $N(i,t)$ como el número de objetos recogidos por el agente i en el ciclo de recogida t.

2. Rendimiento individual esperado

   Se define $\hat{N}(i,t)$ como el número de objetos que el agente i espera recoger en ciclo t ($t > 1$). Su cálculo se realiza a partir de una traza de los rendimientos individuales de los tiempos anteriores.

$$\hat{N}(i,t)=\lambda \hat{N}(i,t-1)+(1-\lambda)\cdot N(i,t-1) \qquad (1)$$

   donde λ es una constante tal que $0 \leq \lambda \leq 1$, que pondera la influencia relativa de los rendimientos pasados en el cómputo del actual. $\hat{N}(i,1)$ se inicializa con el valor de $N(i,1)$.

3. Señal de refuerzo.

   Se define $r(i,t)$ como la señal de refuerzo recibida por el agente i en el ciclo t. El signo de esta señal lo proporciona la diferencia entre el rendimiento actual y el esperado. La división por la traza $\hat{N}(i,t)$, hace que la señal de refuerzo utilizada sea independiente del ambiente donde se desarrolla el aprendizaje pudiendo usarse en diferentes situaciones sin necesidad de variar la constante de aprendizaje.

$$r(i,t)=(N(i,t)-\hat{N}(i,t)) \;/\; \hat{N}(i,t). \qquad (2)$$

4. Tendencia a un comportamiento

   Para todo comportamiento $y \in$ { egoísta, recíproco, altruista }[13], la tendencia que el agente i tiene en el ciclo t a mostrar cada uno de ellos se define como $s_y(i,t)$. Las variables s_y pueden tomar valores reales en el intervalo $[0,\ k_s]$. $s_y(i,0)$ tomará distintos valores según el experimento de que se trata. En cada uno de ellos se especificará cual es el valor de inicialización elegido.

5. Comportamiento mostrado

   El agente i determina el comportamiento c que mostrará en el ciclo t según la siguiente expresión.

$$c(i,t)=y,\quad s_y(i,t)=\max_j[s_j(i,t)+\eta(i,j,t)],\quad j\in\{e,r,a\} \qquad (3)$$

   donde η es una variable aleatoria con distribución uniforme que toma valores enteros en el intervalo $[0,\ k_s\cdot v(i,t)]$. El factor de exploración v depende de la rentabilidad que ha obtenido con anteriores exploraciones.

[13] Por simplicidad, en ocasiones nos referiremos a estas alternativas del comportamientos cooperativo con sus iniciales {e,r,a}.

6. Comportamiento exploratorio.

   Un agente realiza un comportamiento exploratorio si el comportamiento mostrado en ese ciclo, $c(i,t)= z$, no se corresponde con la mayor de las variables de tendencia, esto es, si $s_z \neq \max_j[s_j(i,t)]$.

7. Factor de exploración.

   El factor de exploración v varía en función de la rentabilidad de los ciclos exploratorios. Si un ciclo exploratorio se sigue de una señal de refuerzo positiva, el agente incrementa su tendencia a realizar nuevos ciclos exploratorios incrementando su factor de exploración v.

$$z(i,t) = \begin{cases} \beta\, v(i,t-1) + \delta\, r(i,t-1) & \text{si} \quad c(i,t-1) = y, \quad s_y(i,t-1) \neq \max_j[s_j(i,t-1)],\ j \in \{e,r,a\}, \\ & \text{y} \quad r(i,t-1) > 0 \\ \beta\, v(i,t-1) & \text{en otro caso} \end{cases}$$

$$v(i,t) = \max[z(i,t),\ v_{min}]$$

Mediante esta forma de control de los ciclos exploratorios se consigue una alta exploración al principio del aprendizaje decreciendo ésta a medida que el agente aprende. De esta forma, el rendimiento global del sistema no se ve afectado por efecto de un constante alto grado de exploración. Si el ambiente cambia, el procedimiento de exploración permite una rápida reconfiguración de las tendencias de los agentes, estabilizándose de nuevo una vez aprendida la nueva configuración óptima. En los experimentos del capítulo 5 se muestra este efecto de flexibilidad.

Con las definiciones anteriores, el aprendizaje debe encontrar, para un ambiente determinado, cual es la distribución de comportamientos cooperativos entre los agentes, óptima y estable. Esta distribución de comportamientos será óptima (o cercana al óptimo) si maximiza el número de objetos recogidos por el colectivo. La distribución será estable si todos los agentes tienen estabilizado su comportamiento cooperativo, esto es, para todo i, la probabilidad de que $c(i,t)=y$, es próxima a 1, esto es, el valor de su variable de tendencia $s_y(i)$ es próximo a k_s y el del resto de variables $s_j(i)$ es próximo a cero. La distribución de comportamientos es evolutivamente estable si un agente no mejora su

rendimiento promedio si cambiara de comportamiento cooperativo[14]. Formalmente, sea Y una población de agentes con comportamiento y, y sea $E(y,Y)$ el rendimiento obtenido por el comportamiento y en esa población Y. Sea $E(z,P_{q,z,y})$ el rendimiento obtenido por el comportamiento z en una población P con q agentes con comportamiento z y $(1-q)$ agentes con comportamiento y $(z\neq y)$ En los mismos términos que los planteados en la sección 2.5, el comportamiento y es evolutivamente estable si para todo $(z\neq y)$ se tiene:

$E(y,Y)>E(z,Y)$ o bien

$E(y,Y)=E(z,Y)$ y para una proporción pequeña q
$E(z,P_{q,z,y})< E(y,P_{q,z,y})$.

Al comienzo de un ciclo, el agente decide qué comportamiento cooperativo mostrar de acuerdo a la expresión (3) y calcula, al final de ese ciclo de recogida, su señal de refuerzo (expresiones (1) y (2)). El aprendizaje modifica la tendencia que el agente tiene a mostrar las distintas alternativas del comportamiento cooperativo. El mecanismo es el siguiente: si el agente i en el ciclo t mostró el comportamiento cooperativo y, esto es $c(i,t)=y$, entonces, para $t > 1$ modifica su tendencia a mostrar ese comportamiento,

$$s_y(i,t+1)=s_y(i,t)+\alpha\ r(i,t), \qquad (4)$$

donde α es la tasa de aprendizaje. Finalmente las distintas variables de tendencia de un agente son normalizadas de forma que $\sum_{j\in\{e,r,a\}} s_j(i,t) = k_s$.

Pese al problema de asignación de crédito al que se enfrenta el algoritmo de aprendizaje y la dificultad inicial de la instauración del altruismo, la aplicación de este algoritmo en un colectivo de agentes autónomos demuestra ser eficaz y conduce a la estabilización de estrategias cooperativas óptimas para todos los ambientes en que se han desarrollado los experimentos. Una vez alcanzada la situación de estabilidad, si el ambiente cambia, el algoritmo es lo suficientemente flexible para permitir una reversión de comportamientos hacia una nueva situación de equilibrio que resulte en rendimiento óptimo. Este algoritmo tiene una serie de características de clara inspiración biológica que lo aproximan a las reglas de aprendizaje mencionadas en la sección 2.8. En particular,

- la tendencia de los agentes a mostrar las distintas alternativas de comportamiento se modifica de forma proporcional al beneficio obtenido por las mismas. Si un comportamiento se sigue de una señal de refuerzo positiva, la tendencia a mostrar en el futuro dicho comportamiento se ve incrementada y decrece en caso contrario.

[14] Este último criterio es similar a lo que se conoce como un equilibrio de Nash.

- La inclusión del factor de descuento λ en el cálculo de la traza, hace que los resultados obtenidos por acciones recientes tengan un efecto mayor sobre el comportamiento que acciones anteriores en el tiempo.
- El factor de exploración variable permite a los agentes incrementar su nivel de exploración cuando los ciclos exploratorios previos han resultado beneficiosos y reducirlo progresivamente a un nivel basal (v_{min}) cuando el rendimiento se va maximizando y la exploración no se sigue de resultados beneficiosos. Esto permite que si el ambiente cambia y requiere de una nueva configuración de comportamientos, los agentes abandonen la situación de equilibrio alcanzada. El factor de exploración implementado en este algoritmo es de mayor elaboración que los presentados en los ejemplos de la sección 2.8 y es consistente con observaciones en la naturaleza en las que se aprecia que el comportamiento exploratorio de los animales es dependiente de los beneficios recibidos (Milinski, 1987).

Tabla 4.2. Parámetros usados en las simulaciones.

Nombre	Valor
k_{t_limite}	10000
k_{obj}	80
k_{rec}	5
k_{com}	150
k_m	1
t_{lim}	2000
t_{tft}	2000
Diámetro del robot	7
Rango percepción obstáculos	200
α	50
λ	0.95
β	1
δ	0.95
v_{min}	1
k_s	500

4.3.2 Evaluación del algoritmo de aprendizaje.

Los aspectos a evaluar del efecto del algoritmo de aprendizaje sobre el sistema multiagente son, por un lado, los cambios producidos en el rendimiento global, y por otro lado la configuración de comportamientos en el sistema obtenida al final del proceso de aprendizaje. Para estas evaluaciones se proponen las siguientes pruebas estadísticas:

1. Evaluación del rendimiento

   Se realizan medidas del rendimiento global del equipo de trabajo al inicio y al final del aprendizaje. Este rendimiento global se define como

$$R_g = \sum_{i=1}^{n} N(i)$$

   donde n es el número de agentes en el colectivo y $N(i)$ el número de objetos recogidos por el agente i.

   Se utilizan pruebas estadísticas no paramétricas[15] para contrastar la existencia de diferencias entre los rendimientos globales obtenidos en distintas combinaciones de comportamientos. Igualmente se contrasta la existencia de efecto del aprendizaje sobre este rendimiento global comparando los rendimientos al inicio y final de la fase de aprendizaje.

2. Efecto del aprendizaje sobre la estabilidad de la reciprocidad.

   Para analizar el efecto del aprendizaje sobre la selección de comportamientos por los agentes y para el contraste de la estabilidad de los mismos, se realizan estimaciones puntales y se calculan los intervalos de confianza para un nivel del 99% de la proporción de agentes cooperativos tras el proceso de aprendizaje.

4.4 Sumario del capítulo.

En este capítulo se describe el entorno de experimentación utilizado en la tesis, en sus facetas de simulación e implementación física. Se describen los ambientes de simulación y experimentación física, justificando la necesidad del uso conjunto de ambas metodologías de trabajo.

Se presenta la arquitectura AREA para agentes autónomos y se justifican las elecciones realizadas en el diseño de la misma en cuanto a sus dispositivos sensoriales, motores y de comunicación. En su descripción se enumeran los comportamientos en el repertorio de los agentes, en sus vertientes de comportamientos reactivos fijos y comportamientos modificables mediante aprendizaje. En este último caso, se justifica la inclusión de

[15] La elección de las pruebas estadísticas no paramétricas viene motivada por que los datos obtenidos experimentalmente no cumplen con los requisitos de normalidad ni homocedasticidad requeridos para la utilización de los contrastes paramétricos.

comportamientos cooperativos para que el sistema multiagente mejore su rendimiento. Se argumenta sobre la necesidad de inclusión de procedimientos de adaptación al ambiente y los problemas que ésto genera sobre la estabilidad del altruismo. Se propone un algoritmo de aprendizaje por refuerzo como procedimiento de adaptación y la inclusión de las mecanismos necesarios para la utilización de la estrategia de reciprocidad. Se enuncian los requerimientos de estabilidad de la reciprocidad y se propone la inclusión, en la arquitectura de los agentes, del comportamiento de **`reconocimiento-de-señalizadores`**.

Por último, se realiza la descripción del algoritmo de aprendizaje utilizado y una exposición de las medidas de evaluación que se emplearán en la fase de experimentación.

Capítulo 5

Aprendizaje y estabilización de estrategias altruistas

En este capítulo se presentan los resultados de los diferentes experimentos realizados para confirmar la inestabilidad del altruismo y la eficacia de la estrategia recíproca. A tal fin, se desarrollan experimentos sin el comportamiento recíproco viéndose como el egoísmo invade la población y experimentos que incluyen las tres alternativas (egoísmo, altruismo y reciprocidad) para, mediante el aprendizaje, obtener resultados cercanos al óptimo y estables. Finalmente se muestran experimentos para probar la escalabilidad del sistema y el efecto de la competencia entre agentes. El planteamiento de todos los experimentos es similar. Se tiene un colectivo de 4 agentes desarrollando un tarea de recogida de objetos. Esta tarea se realizará en dos ambientes distintos que se diferencian en la distribución de los objetos en el mundo de trabajo (Tabla 5.1).

En el ambiente A, los objetos son difícilmente accesibles al encontrarse agrupados y rodeados de obstáculos. Por el contrario, el ambiente B presenta los objetos fácilmente accesibles al estar éstos distribuidos uniformemente por el ambiente de trabajo. Los ambientes C y D son mixtos, presentan una distribución de objetos inicial de un tipo y, transcurrido un tiempo, el ambiente cambia presentando los objetos en la distribución contraria. Concretamente, en el ambiente C los agentes recolectan objetos agrupados (igual que en el ambiente A) y tras 900 ciclos de aprendizaje el ambiente cambia para presentar

los objetos distribuidos (ambiente B). En el ambiente D, el sentido de la simulación es el inverso (objetos inicialmente distribuidos y posterior cambio hacia agrupados).

El objetivo del aprendizaje es encontrar, para cada ambiente, una configuración de comportamientos cooperativos en el equipo de trabajo que sea estable y que proporcione al sistema rendimientos globales cercanos al óptimo. Adicionalmente se pretende que la estabilización de estos comportamientos sea reversible si el ambiente de trabajo cambia.

Tabla 5.1. Descripción de los ambientes experimentales

Ambiente	Distribución de Objetos
A	Agrupados
B	Distribuidos
C	Agrupados -> Distribuidos
D	Distribuidos -> Agrupados

5.1 Pruebas sin aprendizaje.

Antes de proceder a los experimentos de aprendizaje, es interesante realizar una caracterización de los ambientes A y B para determinar cual es la distribución de comportamientos cooperativos óptima para cada uno de ellos. Para ello se realizan 100 simulaciones sin aprendizaje de todas las posibles combinaciones del comportamiento cooperativo en los ambientes A y B. Los resultados se muestran en la Tabla 5.2.

A la vista de la Tabla 5.2 se observa que en ambos ambientes existen diferencias significativas en el rendimiento global según la configuración de comportamientos en el sistema (Prueba de Kruskal-Wallis $\chi^2 = 503{,}38$, $p<0{,}01$ para el ambiente A y $\chi^2 = 704{,}55$, $p<0{,}01$ para el ambiente B). En el ambiente A (objetos agrupados), los mejores rendimientos se obtienen cuando todos los agentes son cooperativos (ya sean altruistas o recíprocos). El rendimiento individual de los agentes altruistas decrece a medida que aumenta el número de egoístas en el colectivo. Esto es debido a que el coste de cooperar se distribuye entre menor número de agentes cooperativos (mayor es el tiempo entre relevos). El rendimiento de los agentes egoístas, sin embargo, crece paralelo al incremento de altruistas en la población, no siendo así cuando lo que se incrementa es el número de recíprocos. Estos últimos muestran la estrategia TFT, cooperando en el primer encuentro y "castigando" posteriormente a los agentes no cooperativos. Se evita de esta manera el

constante beneficio que obtienen los agentes no cooperativos de sus abusos hacia los agentes altruistas. Es interesante observar que el rendimiento de altruistas y recíprocos es muy similar en todas las combinaciones en las que coexisten. El rendimiento de un agente cooperativo (altruista o recíproco) sólo depende del número de ellos en la población (dado que afecta a la distribución de costes de permanecer parado señalizando, es decir, al tiempo entre relevos). Esta similitud en rendimientos dificulta el aprendizaje pues la estabilización de la cooperación, como veremos, necesita de la sustitución de altruistas por recíprocos y a igualdad de rendimientos individuales, esta sustitución se ve dificultada.

En el ambiente B (objetos uniformemente distribuidos), la mejor alternativa de comportamiento cooperativo, tanto individual como colectiva es el egoísmo. Si un agente percibe un objeto, la mejor acción posible es recogerlo en lugar de detenerse a señalizar. De hecho, el rendimiento individual de los agentes egoístas es superior al resto en todas las combinaciones posibles de comportamientos y el rendimiento colectivo crece cuando lo hace el número de agentes cooperativos. La presencia innecesaria de cooperación en el equipo de trabajo se traduce en una pérdida de rendimiento global. En esta situación, la ventaja que un agente obtiene por ser ayudado es menor que el coste que supone ayudar, con lo que la cooperación deja de ser ventajosa. Es importante recordar que para la estabilización de la estrategia de reciprocidad se requiere la condición de que el coste de la cooperación sea inferior al beneficio de la ayuda (sección 2.7), condición que no se verifica en el caso del ambiente B. La mejor alternativa del comportamiento cooperativo es entonces la egoísta que, una vez adoptada por el colectivo de agentes, es estable y proporciona rendimientos globales cercanos al óptimo.

Tabla 5.2. Resultados de 100 simulaciones sin aprendizaje en los ambientes A y B. Se presentan las medias y desviaciones típicas de los rendimientos individuales (R_a, R_e y R_r que corresponden a altruista, egoísta y recíproco respectivamente) y global (R_g) para cada combinación posible de los tres comportamientos en el equipo de 4 agentes. Con líneas de puntos se representa la ausencia de un comportamiento en la combinación.

Ambiente A

	R_g		R_a		R_e		R_r	
	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s
AAAA	57,02	8,01	14,26	2,00	----	----	----	----
AAAR	57,28	5,60	14,32	1,48	----	----	14,32	1,87
AAAE	56,00	6,42	12,42	1,42	18,74	3,05	----	----
AARR	57,28	6,60	14,57	1,66	----	----	14,07	1,92
AARE	51,68	8,12	12,00	2,07	15,42	3,02	12,26	2,20
AAEE	50,34	9,85	8,10	1,89	17,07	3,36	----	----
ARRR	56,72	8,76	14,24	2,49	----	----	14,16	2,23
ARRE	47,86	7,95	11,98	2,32	11,86	2,93	12,01	2,26
AREE	40,82	9,09	8,44	2,66	11,81	2,42	8,76	2,35
AEEE	34,14	15,72	0,00	0,00	11,38	5,24	----	----
RRRR	57,06	7,82	----	----	----	----	14,27	1,96
RRRE	44,40	8,39	----	----	8,14	2,68	12,09	2,46
RREE	28,44	6,00	----	----	5,81	1,35	8,41	2,28
REEE	12,78	2,72	----	----	3,67	0,77	1,78	1,33
EEEE	10,04	2,87	----	----	2,51	0,72	----	----

Ambiente B

	R_g		R_a		R_e		R_r	
	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s
AAAA	36,20	7,28	9,05	1,82	----	----	----	----
AAAR	35,32	7,52	9,01	2,07	----	----	8,30	2,19
AAAE	63,00	3,59	2,23	1,10	56,32	1,80	----	----
AARR	36,30	7,48	9,20	2,00	----	----	8,95	2,08
AARE	66,90	4,33	3,77	1,47	54,24	1,81	5,12	1,15
AAEE	111,12	3,49	0,67	0,48	54,89	1,67	----	----
ARRR	35,44	9,62	9,10	2,61	----	----	8,78	2,47
ARRE	68,82	3,77	4,64	1,72	52,90	1,79	5,64	1,20
AREE	111,92	3,46	2,48	1,27	52,72	1,50	4,00	0,64
AEEE	158,98	3,16	0,00	0,00	52,99	1,05	----	----
RRRR	35,66	8,01	----	----	----	----	8,92	2,00
RRRE	71,84	5,30	----	----	51,68	2,18	6,72	1,44
RREE	111,66	2,92	----	----	51,37	1,46	4,46	0,46
REEE	188,84	5,06	----	----	50,91	1,35	36,12	3,07
EEEE	202,92	4,59	----	----	50,73	1,15	----	----

A continuación se presentan los cinco tipos de experimentos de aprendizaje realizados con una descripción de los objetivos perseguidos por cada uno de ellos.

1. Inestabilidad del altruismo puro.

   El objetivo de este experimento es demostrar el carácter inestable del altruismo cuando se enfrenta a la alternativa egoísta, es decir, demostrar experimentalmente que el altruismo no es una *estrategia evolutivamente estable* (ESS) (Maynard-Smith, 1982). El experimento se realiza en el ambiente A y se enfrentan las alternativas de comportamiento *altruista* y *egoísta* en ausencia de la alternativa de reciprocidad.

2. Estabilización del altruismo mediante la estrategia de reciprocidad

   Estos experimentos pretenden demostrar que la inclusión de la estrategia de reciprocidad en el repertorio de comportamientos de los agentes estabiliza la cooperación altruista frente a la invasión de la estrategia egoísta. En este caso se enfrentan las tres posibles alternativas de comportamiento cooperativo (egoísta, recíproca y altruista) en los ambientes A y B. El experimento se realiza en ambos ambientes con la intención de probar que la emergencia del altruismo recíproco es dependiente del ambiente. Esto es, la reciprocidad es una solución para la estabilidad del altruismo, seleccionable mediante aprendizaje. Su inclusión en el repertorio de comportamientos no fuerza a los agentes a cooperar en aquellos ambientes donde no sea ventajoso.

3. Flexibilidad

   Para mostrar la flexibilidad del algoritmo de aprendizaje, se desarrollan los experimentos del tipo 2 en los ambientes C y D. Se pretende demostrar que, una vez estabilizada una estrategia de comportamiento adecuada para un ambiente, si éste cambia, el algoritmo de aprendizaje permite a los agentes modificar sus comportamientos hasta que el sistema alcanza una nueva configuración de comportamientos óptima y estable para las nuevas condiciones.

4. Escalabilidad

   Mediante este serie de experimentos se pretende demostrar que las conclusiones obtenidas de los experimentos 2 y 3 son válidas para colectivos de agentes de distinto tamaño. Se repetirán los experimentos de las series 2 y 3 con distinto número de agentes en el sistema multiagente.

5. Efecto de la competencia.

   Los experimentos en las series 1, 2 y 3 se realizan en condiciones de trabajo no limitantes. Esto es, los objetos en el mundo de trabajo se encuentran en exceso (inagotables). Cambios en esta condición se traducen en la aparición de efectos de competencia por los recursos que serán analizados en estos experimentos.

5.2 Inestabilidad del altruismo

En este primer experimento, los agentes sólo pueden mostrar las alternativas de comportamiento cooperativo altruista y egoísta. Se inicializa sus variables de tendencia de comportamiento de forma que al principio del aprendizaje muestran todos el comportamiento altruista ($s_a(i,0)=k_s$ y $s_e(i,0)=0$, $\forall i$, $n=4$). El experimento se desarrolla en el ambiente A.

La Figura 5.1 muestra la evolución del rendimiento colectivo y el número de agentes mostrando cada una de las dos alternativas de comportamiento cooperativo a lo largo de los ciclos de aprendizaje. En esta figura puede observarse que existe una rápida caída del número de agentes altruistas, que son sustituidos por los egoístas. En muy pocos ciclos de aprendizaje, algunos agentes comienzan a seleccionar la alternativa egoísta, obteniendo un mayor rendimiento individual en perjuicio del rendimiento global. Sin embargo, cuando la invasión del egoísmo es completa, se observa una pérdida en el rendimiento individual de todos los agentes y una mayor disminución en el rendimiento colectivo del sistema. Al final del aprendizaje, el rendimiento global es menor del 50 % del que se obtenía al inicio cuando todos los agentes eran altruistas. En la Tabla 5.3 se muestran los resultados numéricos del experimento. Un test de rangos con signo (Wilcoxon) muestra las diferencias altamente significativas en el rendimiento global obtenido antes y después del proceso de aprendizaje. Asimismo se muestra el promedio de agentes en cada alternativa de comportamientos y el intervalo de confianza al 99% de la proporción de altruistas en el colectivo de agentes al final del aprendizaje.

Tabla 5.3. Valores medios y desviaciones típicas sobre 100 repeticiones del rendimiento global (R_g) al inicio y final de la fase de aprendizaje y número de agentes mostrando cada alternativa de comportamiento al final del aprendizaje.

	Rendimiento Global		Nº medio de agentes por comportamiento	
	$\bar{x}$	s	Egoísta	Altruista
Inicial	56,75	7,62	3,75	0,25
Final	16,41	13,15		
	Z = -8,65	p<0,001	IC(99%) $p_{altruista}$ = 0,062 ±0,031	

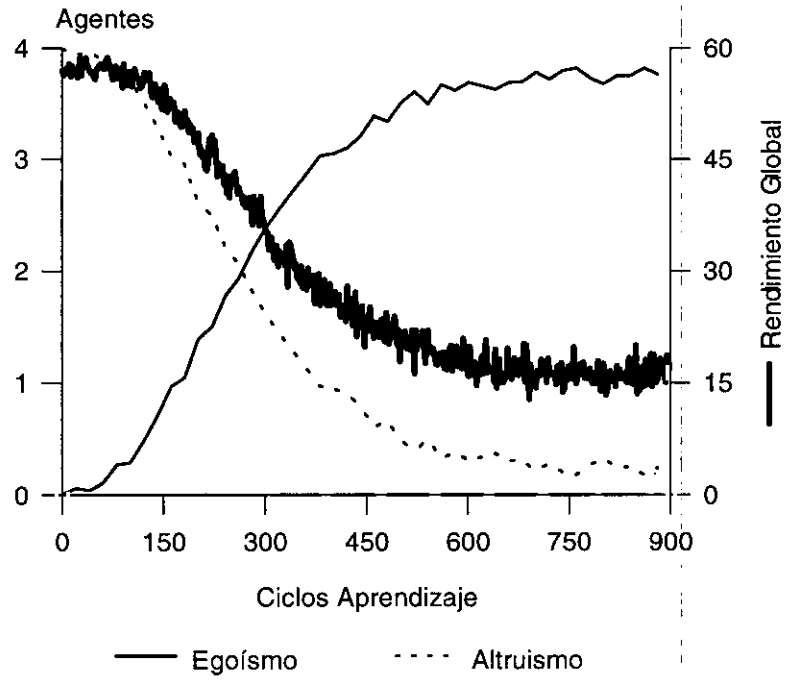


Figura 5.1. Resultados medios de 100 repeticiones del experimento de aprendizaje en el ambiente A en ausencia del comportamiento recíproco. Se muestra el rendimiento global del sistema y el número de agentes exhibiendo cada una de las dos alternativas de comportamiento.

Los resultados demuestran el carácter inestable del altruismo, que es sustituido por el egoísmo, así como el paradójico comportamiento del rendimiento individual de los agentes con comportamiento egoísta. Pese a que el óptimo global se encuentra cuando todos los agentes cooperan, la "tentación" de no cooperar, es decir la diferencia entre el rendimiento que obtienen los altruistas y egoístas en cualquiera de las combinaciones (Tabla 5.2), hace que sea irremediable esa invasión del egoísmo. Esta invasión hace que el rendimiento individual de los agentes acabe siendo menor que si afrontaran los gastos de la cooperación. Los agentes sucumben a la tentación del egoísmo pues supone una mejora en el rendimiento individual a corto plazo. No obstante, esta decisión provoca un deterioro en el rendimiento final de esos mismos agentes. De acuerdo al criterio de estabilidad evolutiva expuesto en el capítulo 2, si el rendimiento de un agente egoísta en una población con individuos altruistas (por ejemplo AAAE) es superior al que obtiene un altruista en una población formada por altruistas (AAAA), bajo las leyes de la selección natural (el aprendizaje en nuestro caso), el egoísmo invadirá la población. Esto determina

que el altruismo no es una estrategia evolutivamente estable. Esta condición es generalizable a toda población en la que coexistan individuos con comportamientos altruistas y egoístas. La Tabla 5.2 y la Figura 5.1 confirman este resultado.

5.3 Estabilización del altruismo mediante la estrategia TIT FOR TAT

En este experimento, los agentes pueden mostrar las tres alternativas de comportamiento cooperativo (egoísta, recíproco y altruista). La inclusión de la estrategia de reciprocidad permite que los agentes muestren la estrategia TFT, cooperando en el primer encuentro y repitiendo posteriormente el comportamiento cooperativo mostrado por su contrincante. Esto hace que los agentes tengan una posibilidad de evitar ser explotados pues responden "castigando" al individuo no cooperativo ignorando sus peticiones de ayuda. Los agentes inicializan sus variables de tendencia de comportamiento de forma que al principio del aprendizaje muestran todos el comportamiento altruista ($s_a(i,0)=k_s$, $s_r(i,0)=0$ y $s_e(i,0)=0$, $\forall i$, $n=4$). El experimento se desarrolla en los ambientes A y B.

La Figura 5.2 muestra la evolución del rendimiento colectivo y el número de agentes mostrando cada una de las alternativas de comportamiento cooperativo según procede el aprendizaje. En la Tabla 5.4 y Tabla 5.5 se muestran los valores medios y desviaciones típicas del rendimiento global así como el número de agentes mostrando cada comportamiento en la fase final del aprendizaje, calculados sobre 100 repeticiones, en los ambientes A y B respectivamente. Se muestra el promedio de agentes mostrando cada alternativa de comportamiento cooperativo y el intervalo de confianza para la proporción de agentes cooperativos en el colectivo al final del aprendizaje. En el ambiente A (Tabla 5.4), la prueba de rangos con signos (Wilcoxon) no muestra diferencias significativas en el rendimiento global por efecto del aprendizaje, es decir, la sustitución del altruismo por la estrategia de reciprocidad mantiene el rendimiento del sistema cercano al óptimo. Por el contrario, en el ambiente B (Tabla 5.5), el análisis estadístico demuestra un incremento significativo del rendimiento global al final del aprendizaje. En las primeras fases del mismo, el comportamiento de los agentes altruistas es peor y su sustitución por agentes egoístas se traduce en una mejora del rendimiento estadísticamente significativa. La comparación de la estimación por intervalos de las proporciones de agentes recíprocos en ambos ambientes ilustra el efecto del ambiente sobre la afinidad de los agentes a mostrar el comportamiento recíproco.

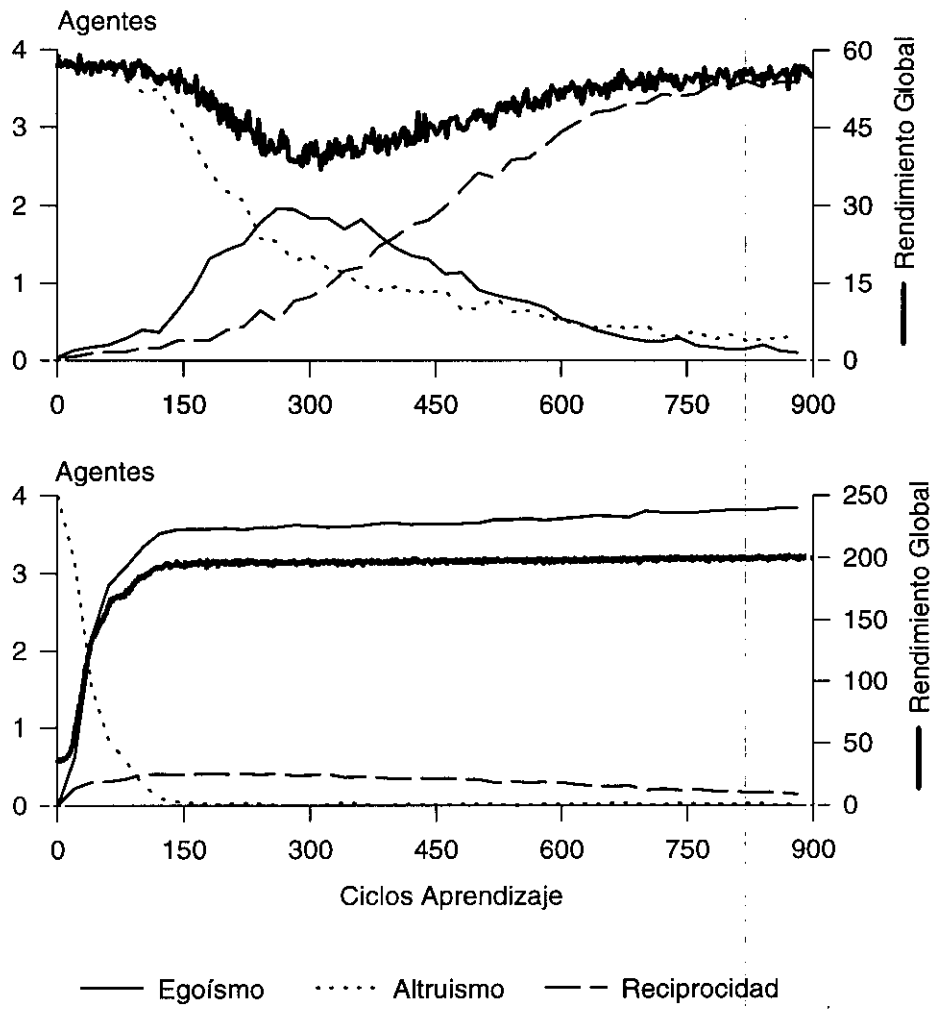


Figura 5.2. Resultados medios de 100 repeticiones de los experimentos de aprendizaje en el ambiente A (panel superior) y en el ambiente B (panel inferior). Se muestra el rendimiento global del sistema y el número de agentes exhibiendo cada una de las alternativas de comportamiento cooperativo (egoísta, recíproca y altruista).

Tabla 5.4. Valores medios y desviaciones típicas sobre 100 repeticiones del rendimiento global (R_g) al inicio y final de la fase de aprendizaje y número de agentes mostrando cada alternativa de comportamiento al final del aprendizaje.

	Ambiente A				
	Rendimiento Global		N° medio de agentes por comportamiento		
	$\overline{x}$	s	Egoísta	Recíproco	Altruista
Inicial	56,75	7,51	0,09	3,66	0,25
Final	55,74	11,52			
	Z =-0,17	p=0,865	IC(99%) $p_{cooperativos}$ = 0,977 ± 0,019		

Tabla 5.5. Valores medios y desviaciones típicas sobre 100 repeticiones del rendimiento global (R_g) al inicio y final de la fase de aprendizaje y número de agentes mostrando cada alternativa de comportamiento al final del aprendizaje.

	Ambiente B				
	Rendimiento Global		N° medio de agentes por comportamiento		
	$\overline{x}$	s	Egoísta	Recíproco	Altruista
Inicial	35,65	7,68	3,85	0,10	0,05
Final	201,40	6,38			
	Z = -8,68	p<0,01	IC(99%) $p_{cooperativos}$ = 0,037 ± 0,024		

En la Figura 5.2 se observa que el colectivo de agentes alcanza, en ambos ambientes, la distribución óptima de comportamientos cooperativos. Para el ambiente A (panel superior), aunque existen varias configuraciones con rendimientos cercanos al óptimo, sólo aparecen al final del aprendizaje aquellas que son estables (RRRR y ARRR), es decir las únicas en las que se cumple que ningún agente puede mejorar su rendimiento individual al cambiar de comportamiento cooperativo. Si la combinación de comportamientos contuviera agentes altruistas en exceso, éstos podrían ser explotados por los agentes egoístas, produciéndose entonces la invasión de la población. En la primera parte de la gráfica (Figura 5.2 - superior), se observa que los agente egoístas comienzan a invadir al colectivo inicializado en altruismo, provocando una pérdida de rendimiento global apreciable. Esta invasión es completa aunque no puede apreciarse en su totalidad en la gráfica porque ésta representa

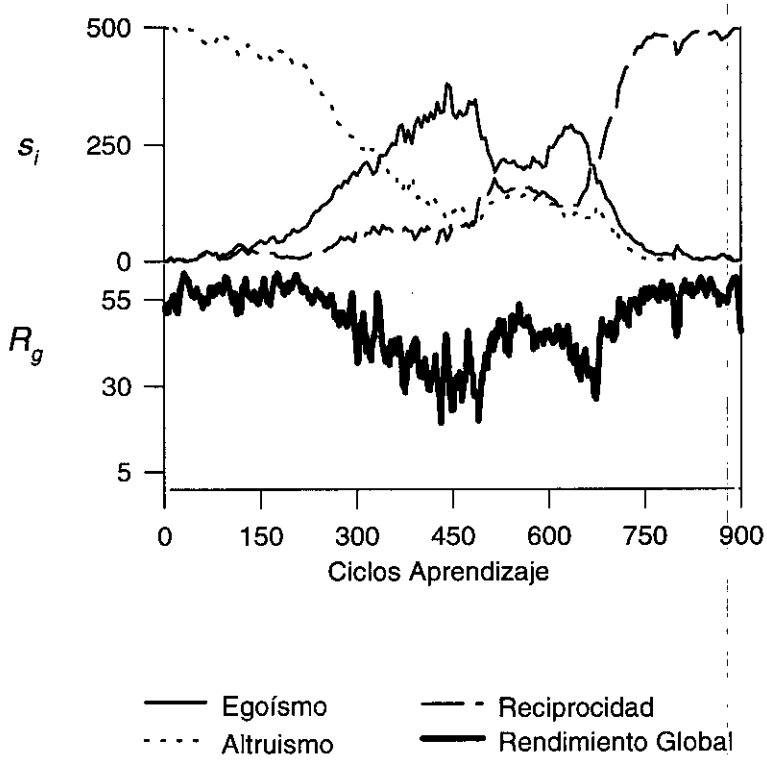


Figura 5.3. Resultados típicos de una simulación aislada del experimento en el ambiente A. Se muestra el valor de las variables s_i de tendencia a los distintos comportamientos a lo largo del aprendizaje.

los valores de 100 repeticiones y en cada una de ellas la invasión sucede en distintos tiempos amortiguándose las dimensiones del efecto en la gráfica. Para ilustrar mejor el proceso, en la Figura 5.3 se muestran los resultados de una simulación aislada. En el momento de la invasión, el rendimiento global decrece. A partir de ese punto comienza a recuperarse la cooperación gracias a la aparición de los individuos recíprocos. Al final del aprendizaje, la recuperación del rendimiento global es completa, alcanzando el sistema una configuración de comportamientos cercana al óptimo y estable frente a nuevas invasiones.

Es necesario hacer una reflexión acerca de los factores que dificultan el aprendizaje en el ambiente A. En primer lugar, en la Tabla 5.2 puede observarse que el rendimiento de altruistas y recíprocos es muy similar en muchas combinaciones con lo que es fácil que ambos comportamientos coexistan en la población durante alguna fase del aprendizaje. Sin embargo, la presencia de individuos altruistas en el equipo facilita la posterior aparición de agentes egoístas dado que éstos últimos mejoran mucho su rendimiento por la ayuda incondicional que reciben de los primeros. Una segunda dificultad surge del ocasional abuso que los agentes egoístas pueden ejercer sobre los agentes recíprocos. Por un lado, el

tipo de comunicación utilizada, tipo *broadcast* (no dirigida hacia ningún agente en particular), hace que los mensajes de donación emitidos por un agente recíproco hacia otro cooperativo pueden ser percibidos por cualquier agente dentro de un rango k_{com}. De esta forma, se tiene que, ocasionalmente, un agente egoísta se aprovecha de mensajes de ayuda emitidos hacia agentes cooperativos. Por otro lado, los agentes egoístas se benefician de la estrategia TFT pues ésta implica una cooperación inicial indiscriminada hacia cualquier agente extraño. Ambos factores explican el mejor rendimiento que presentan, por ejemplo, los egoístas en una combinación RREE frente a una combinación EEEE. Por último, la recuperación de la cooperación en el equipo tras la invasión egoísta requiere de la aparición simultánea de al menos 2 agentes cooperativos para así poder notar los efectos ventajosos de la cooperación. La aparición de forma aislada, por el contrario, se traduce en pérdidas en el rendimiento individual con lo que los agentes disminuyen su tendencia a cooperar en el futuro. Este último aspecto negativo se disminuye mediante la inclusión en el algoritmo de aprendizaje del factor de exploración variable que, ante ciclos exploratorios ventajosos, incrementa la probabilidad de seguir explorando, disminuyendo ésta a medida que el agente estabiliza su rendimiento (ver sección 4.3.1). Pese a todas estas dificultades, el algoritmo de aprendizaje consigue estabilizar una distribución óptima de comportamientos.

En el ambiente B (Figura 5.2 abajo), la configuración óptima no requiere la presencia comportamientos recíprocos ni altruistas en la combinación del equipo de trabajo. El mejor comportamiento posible en ese ambiente es el egoísta. Partiendo de una configuración totalmente altruista, se observa un rendimiento por debajo del óptimo. A medida que avanza el aprendizaje, los agentes eligen la alternativa egoísta que es individual y colectivamente más rentable. Al final del aprendizaje se alcanza la configuración óptima y estable (EEEE). La velocidad de aprendizaje es superior en este caso debido a que, por un lado, los agentes egoístas sistemáticamente obtienen mejores rendimientos que el resto de los comportamientos (Tabla 5.2) y, a diferencia del aprendizaje en el ambiente A, la recuperación de la estrategia óptima no necesita de la aparición simultánea de un mínimo de agentes con el comportamiento óptimo, sino que es suficiente catalizar la recuperación del sistema mediante la aparición de agentes egoístas aislados.

Como conclusión de esta serie de experimentos observamos que el algoritmo de aprendizaje propuesto, implementado en la arquitectura AREA, consigue seleccionar y estabilizar distribuciones óptimas de comportamientos cooperativos en ambientes distintos.

5.4 Adaptabilidad del sistema frente a cambios en el ambiente de trabajo

Esta serie de experimentos está destinada a mostrar la capacidad de sistema multiagente para revertir sus comportamientos ante cambios ambientales. Los agentes pueden mostrar las tres alternativas de comportamiento cooperativo (egoísta, recíproco y altruista). Los agentes inicializan sus variables de tendencia de comportamiento de forma que al principio del aprendizaje muestran todos el comportamiento altruista ($s_a(i,0)=k_s$, $s_r(i,0)=0$ y $s_e(i,0)=0$, $\forall i$, $n=4$). El experimento se desarrolla en los ambientes C y D, esto es, comienza el aprendizaje en un ambiente y transcurridos 900 ciclos de aprendizaje, el ambiente cambia al opuesto (Tabla 5.1).

La Figura 5.4 muestra la evolución del rendimiento colectivo y el número de agentes mostrando cada una de las alternativas de comportamiento cooperativo a lo largo del aprendizaje. En la Tabla 5.6 y Tabla 5.7 se muestran los valores medios y desviaciones típicas del rendimiento global en el momento del cambio de ambiente y final del aprendizaje así como el número medio de agentes mostrando cada alternativa del comportamiento cooperativo al final de la fase de aprendizaje, calculados sobre 100 repeticiones. Las pruebas estadísticas (rangos con signo) muestran diferencias significativas en el rendimiento global del sistema entre el momento inmediatamente posterior al cambio de ambiente y el final del aprendizaje. Se muestra también un intervalo de confianza para la proporción de agentes cooperativos (recíprocos y altruistas) al final del aprendizaje.

Estos resultados muestran la flexibilidad del algoritmo de aprendizaje para enfrentarse a cambios de ambiente. Una vez que el equipo de agentes alcanza la configuración óptima de comportamientos cooperativos para un ambiente determinado, si el ambiente cambia, el equipo se adapta a la nueva situación encontrando de nuevo la configuración óptima y estable. En el ambiente C, los primeros ciclos de aprendizaje se realizan con los objetos agrupados como en el ambiente A. Al igual que en los experimentos del apartado anterior, el equipo encuentra una configuración de comportamientos estable (RRRR o bien ARRR) proporcionando un rendimiento global cercano al óptimo. Tras el cambio de ambiente (se pasa a objetos distribuidos), el rendimiento cae súbitamente dado que la configuración del equipo de trabajo no es adecuada para la nueva situación. Los agentes, a través de los ciclos exploratorios, cambian su comportamiento progresivamente, abandonando la estrategia recíproca y sustituyéndola por la egoísta. Al final del aprendizaje, la combinación es EEEE que es estable y el rendimiento global del equipo alcanza niveles cercanos al óptimo. En el ambiente D, la secuencia de situaciones ambientales es la inversa. De la configuración estabilizada en el ambiente con objetos distribuidos (EEEE), se alcanza una configuración óptima y estable para el nuevo ambiente (RRRR o ARRR).

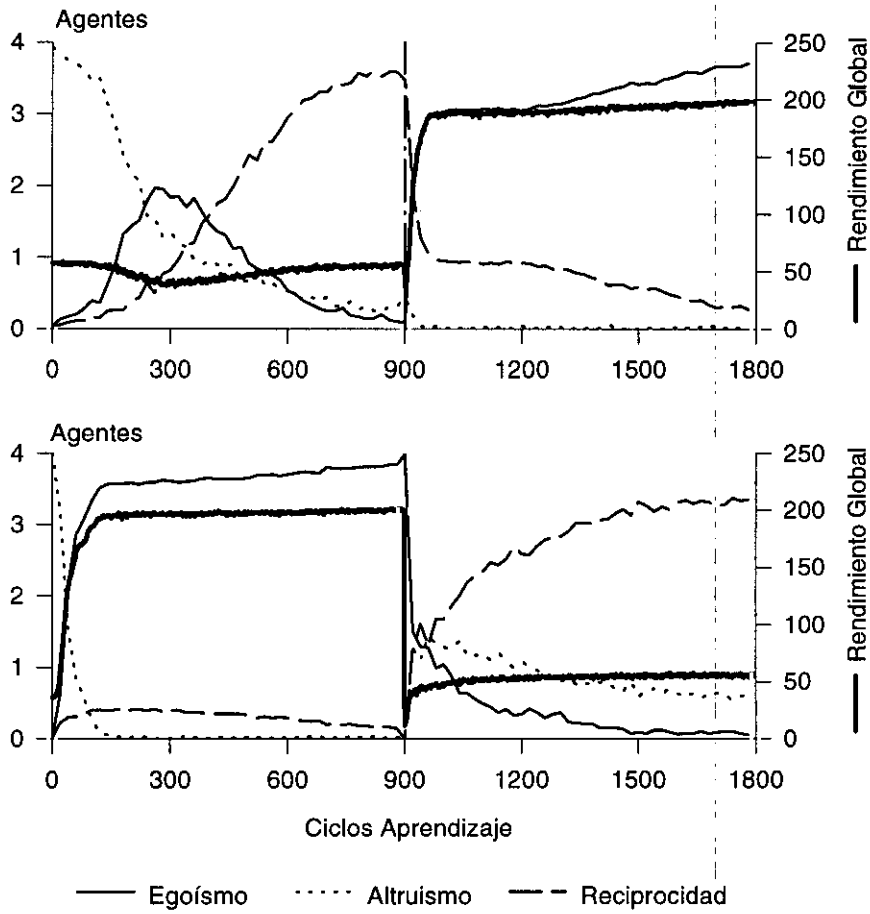


Figura 5.4. Resultados medios de 100 repeticiones de los experimentos de aprendizaje en el ambiente C (arriba) y en el ambiente D (abajo). Se muestra el rendimiento global del sistema y el número de agentes exhibiendo cada una de las alternativas de comportamiento cooperativo (egoísta, recíproca y altruista). El cambio de ambiente se produce en el ciclo 900.

Es importante notar que el punto de partida del aprendizaje tras el cambio de ambiente es distinto al punto de partida en los ambientes originales. En el caso del ambiente C, por ejemplo, el cambio de ambiente se produce cuando todos los agentes tienen su variable s_r en valores cercanos al límite k_s. Ese punto de partida es distinto, cuantitativa y cualitativamente, al que se tenía en los experimentos del apartado 5.2, en los cuales se partía de una situación de saturación completa en altruismo puro. Ambos hechos se traducen en que el aprendizaje es más rápido tras un cambio de ambiente que cuando la inicialización de los comportamientos es en el extremo de altruismo puro. Los resultados obtenidos en los experimentos de flexibilidad demuestran que el aprendizaje tiene lugar

con independencia del punto de inicialización de las variables de comportamiento. Sea cual sea la inicialización, el aprendizaje converge a la distribución de comportamientos más indicada en cada ambiente.

Estos resultados, junto con los obtenidos en la sección 5.3, demuestran que la inclusión de la alternativa de comportamiento recíproca en la arquitectura AREA y la utilización del algoritmo de aprendizaje propuesto, conducen a los agentes a seleccionar y estabilizar la estrategia de comportamientos óptima para cada ambiente. Esta estabilización es reversible si las condiciones ambientales requieren de una nueva configuración de comportamientos.

Tabla 5.6. Valores medios y desviaciones típicas sobre 100 repeticiones del rendimiento global (R_g) en el momento del cambio de ambiente y al final de la fase de aprendizaje y número de agentes mostrando cada alternativa de comportamiento al final del aprendizaje.

	Ambiente C				
	Rendimiento Global		Nº medio de agentes por comportamiento		
	$\overline{x}$	s	Egoísta	Recíproco	Altruista
Cambio	37,47	14,22	3,64	0,30	0,06
Final	197,76	8,96			
	Z = -8,68	p<0,01	IC(99%) $p_{cooperativos}$ = 0,09 ± 0,003		

Tabla 5.7. Valores medios y desviaciones típicas sobre 100 repeticiones del rendimiento global (R_g) en el momento del cambio de ambiente y al final de la fase de aprendizaje y número de agentes mostrando cada alternativa de comportamiento al final del aprendizaje.

	Ambiente D				
	Rendimiento Global		Nº medio de agentes por comportamiento		
	$\overline{x}$	s	Egoísta	Recíproco	Altruista
Cambio	11,32	5,87	0,06	3,35	0,59
Final	56,31	9,23			
	Z = -8,68.	p<0,01	IC(99%) $p_{cooperativos}$ = 0,982 ± 0,016		

5.5 Escalabilidad del sistema multiagente

En este grupo de experimentos se pretende demostrar que el funcionamiento del sistema multiagente propuesto con la arquitectura AREA y el algoritmo de aprendizaje se mantiene en diferentes tamaños de grupo. Para ello se altera el número de agentes en el colectivo. Se usan los mismos ambientes que en los apartados anteriores pero en este caso con un equipo de trabajo formado por 12 agentes. Para estos experimentos los agentes pueden elegir entre las alternativas de comportamiento recíproco y egoísta. No se ha incluido el comportamiento altruista puro dado que el interés del estudio se centra ahora en el efecto del cambio de tamaño del sistema multiagente sobre la estabilidad de las soluciones óptimas encontradas en los ambientes propuestos.

5.5.1 Pruebas sin aprendizaje.

Antes de proceder a los experimentos de aprendizaje, se realiza una caracterización de los ambientes A y B para determinar, para esta nueva situación de 12 agentes, cual es la evolución de los rendimientos individuales y del colectivo según varía el número de agentes recíprocos en el colectivo. La comparación de estos resultados con los obtenidos con un sistema formado por 4 agentes nos informará de los efectos del cambio de escala. Los posteriores experimentos con aprendizaje demuestran que el comportamiento del sistema es equivalente en ambas escalas.

Para la caracterización de los rendimientos individuales y global, se realizan 100 simulaciones sin aprendizaje de todas las posibles distribuciones de los comportamientos cooperativos egoísta y recíproco en los ambientes A y B. Los resultados se muestran en la Figura 5.5 y en la Tabla 5.8. En ambos casos se representan las medias y desviaciones del rendimiento global (R_g) y de los rendimientos individuales de las dos alternativas de comportamiento (R_r y R_e). Una prueba de Kuskal-Wallis muestra diferencias significativas en el rendimiento global según las combinaciones de comportamientos (Prueba de Kruskal-Wallis $\chi^2 = 601{,}67$ $p<0{,}01$ para el ambiente A y $\chi^2 = 572{,}22$ $p<0{,}01$ para el ambiente B). En el apartado relativo al ambiente A de la Tabla 5.8, se observa que el rendimiento global se incrementa a medida que lo hace el número de agentes recíprocos en el colectivo. Este efecto es similar al encontrado en el caso de un sistema con 4 agentes. Sin embargo, en el caso actual se aprecia con mejor detalle que este aumento no es lineal (Figura 5.5-izquierda). Esta no linealidad se debe fundamentalmente a la distinta influencia que el cambio de comportamiento de un agente egoísta hacia recíproco tiene, directamente sobre su rendimiento individual e indirectamente sobre los rendimientos individuales de sus compañeros de trabajo.

Tabla 5.8. Resultados de 100 simulaciones sin aprendizaje en los ambientes A y B. Se presentan las medias y desviaciones típicas de los rendimientos individuales (R_e y R_r que corresponden a egoísta y recíproco respectivamente) y del global (R_g) para las combinaciones posible de ambos comportamientos en el equipo de 12 agentes.

Ambiente A

	R_g		R_r		R_e	
nº recíprocos	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s
0	32,78	4,66	----	----	2,73	0,39
1	44,16	6,83	1,60	1,37	3,87	0,59
2	81,06	9,78	8,48	2,26	6,41	0,68
3	113,24	13,05	12,56	2,03	8,40	0,92
4	138,34	15,57	14,37	1,83	10,11	1,24
5	157,16	17,06	15,26	1,92	11,55	1,24
6	178,34	11,89	16,84	1,09	12,88	1,21
7	190,34	13,94	17,08	1,37	14,16	1,11
8	198,50	14,29	17,31	1,28	15,01	1,54
9	207,60	13,69	17,87	1,26	15,58	1,41
10	212,74	9,70	18,15	0,83	15,60	1,74
11	221,86	11,64	18,61	1,02	17,16	2,52
12	221,92	9,46	18,49	0,79	----	----

Ambiente B

	R_g		R_r		R_e	
nº recíprocos	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s
0	607,92	12,79	----	----	50,66	1,07
1	594,72	9,77	31,76	4,43	51,18	0,87
2	519,88	9,58	4,06	0,47	51,18	0,97
3	477,68	8,95	5,53	1,72	51,23	0,95
4	444,68	10,57	8,40	1,98	51,39	0,87
5	433,08	9,01	13,87	1,25	51,96	0,99
6	437,04	8,05	20,46	0,63	52,38	0,99
7	445,66	8,21	25,83	0,70	52,97	0,94
8	450,88	9,50	29,65	0,78	53,42	1,16
9	457,96	7,47	32,93	0,60	53,87	1,06
10	464,24	7,91	35,55	0,63	54,35	1,59
11	466,64	10,65	37,47	0,90	54,48	1,78
12	472,50	7,99	39,38	0,67	----	----

En un sistema con m agentes de los que n son recíprocos ($m > n$), el cambio de n a n+1 recíprocos supone:

1. Los n agentes recíprocos incrementan su rendimiento individual pues el tiempo que pasan parados señalizando (tiempo entre relevos) es menor, al repartirse ahora entre n+1 agentes, y disminuye el tiempo que tardan en encontrar el primer objeto, que es una variable con ley de probabilidad exponencial que decrece en función de n. La importancia de ambos factores decrece cuando n aumenta.

2. El agente recíproco (n+1)-ésimo incrementa su propio rendimiento individual al pasar de egoísta a recíproco. En la Tabla 5.8 se tiene que $R_r > R_e$ en todas las combinaciones

3. Los agentes egoístas del sistema (m–n–1) también incrementan su rendimiento debido a que utilizarán la cooperación inicial de n+1 agentes que usan la estrategia TFT y se incrementa también la probabilidad de poder aprovecharse de las señales tipo "broadcast" que emiten los agentes recíprocos entre ellos.

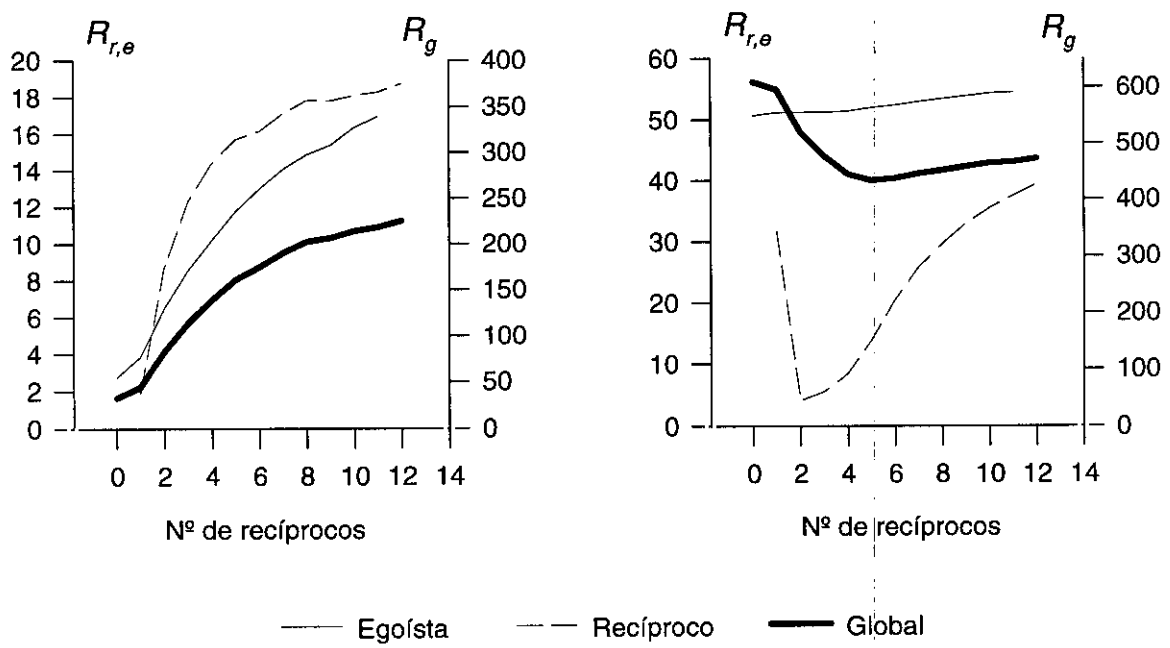


Figura 5.5. Evolución de los rendimientos individuales ($R_{r,e}$) y del rendimiento global (R_g) según el número de agentes recíprocos en el equipo de 12 agentes en el ambiente A (izquierda) y en el ambiente B (derecha). Se representan las medias y desviaciones de dichas medidas en 100 repeticiones sin aprendizaje.

En el análisis de los resultados en el ambiente B (Tabla 5.8 y Figura 5.5 derecha), se observa que los rendimientos individuales y el global tienen una evolución radicalmente distinta a la observada en el ambiente A. La comparación de los rendimientos individuales de cada uno de los comportamientos ofrece que el comportamiento egoísta obtiene siempre mayor rendimiento individual que el recíproco. El rendimiento óptimo se obtiene cuando todos los agentes del colectivo muestran la alternativa de comportamiento egoísta. Este rendimiento decrece a medida que aumenta el número de agentes recíprocos, sin embargo, esta disminución se interrumpe cuando el número de agentes recíprocos es superior a un punto crítico de cuatro agentes. Cuando el equipo de trabajo presenta un número de recíprocos menor de cuatro, cada agente cooperativo adicional se traduce muy frecuentemente en un agente detenido señalizando adicional, con el consiguiente detrimento en el rendimiento global. A partir del quinto agente recíproco, el rendimiento global del sistema se incrementa suavemente debido a que los agentes recíprocos adicionales ya no se bloquean señalizando sino que la mayoría de las veces relevan a alguno de los cuatro anteriores[1]. Así, cuando el equipo de trabajo presenta cinco o más agentes cooperativos, un promedio de 4 están simultáneamente detenidos y el resto relevándoles en su tarea. A partir de este punto, a mayor número de recíprocos, se incrementa su rendimiento individual pues se distribuye entre mayor número de agentes el coste de permanecer detenidos señalizando. El caso de un agente recíproco aislado es un caso particular en el que ese agente obtiene un rendimiento individual mucho mayor que el resto de configuraciones. Esto es debido a que cuando un agente recíproco aislado consume su tiempo límite de señalización sin haber cooperado con ningún agente y sin haber sido relevado, deja de señalizar sin volver a hacerlo durante el resto del ciclo de recogida. Cuando son dos o más agentes cooperativos, éstos cooperan entre sí en los primeros ciclos bloqueándose posteriormente al señalizar objetos distantes que impiden la percepción de la señal de relevo.

5.5.2 Experimentos con aprendizaje.

Para los experimentos de aprendizaje se sigue un protocolo de experimentación similar al del las series de experimentos 5.3 y 5.4. Se diferencia en que el repertorio de comportamientos de los agentes no incluye el comportamiento altruista puro y la inicialización de las variables de tendencia de los comportamientos es dependiente de la distribución de objetos que haya en el momento de iniciarse el aprendizaje, para situar al sistema en la configuración más alejada del óptimo. Si el aprendizaje comienza con lo objetos agrupados (ambientes A y C), los valores iniciales de las variables de

[1] Este promedio de 4 agentes simultáneamente detenidos señalizando viene determinado por las características del ambiente que incluyen la disposición de los obstáculos y el rango de percepción de las señales de relevo.

comportamiento son $s_r(i,0)=0$ y $s_e(i,0)=k_s$ ($\forall i$, $n=4$). Por el contrario, si el aprendizaje se inicia con los objetos distribuidos (ambientes B y D), estos valores serán ($s_r(i,0)=k_s$ y $s_e(i,0)=0$, $\forall i$).

La Figura 5.6 presenta los resultados de aprendizaje del equipo de 12 agentes en los ambientes A, B, C y D. Se observa que en todos ellos se alcanza la configuración óptima y estable de una forma similar a como lo hacía un sistema de 4 agentes. En la Tabla 5.9 se muestran los valores medios y desviaciones típicas del rendimiento global al principio y al final del aprendizaje y el número medio de agentes en cada comportamiento al final del mismo, calculados sobre 100 repeticiones. Se muestra el intervalo de confianza (99%) para la proporción de agentes recíprocos al final del aprendizaje para los 4 ambientes. La amplitud de estas estimaciones en los ambientes A y B demuestran la diferente composición de comportamientos cooperativos en ambos ambientes al final del aprendizaje. Las pruebas estadísticas (prueba de rangos con signo) realizadas sobre los cambios en el rendimiento global ofrece resultados similares a los obtenidos en el sistema de 4 agentes.

Los resultados de los experimentos de aprendizaje con un colectivo de 12 agentes son equivalentes a los obtenidos con un sistema de 4 agentes. En el caso de 12 agentes, se tiene que la diferencia entre los rendimientos individuales de ambos comportamientos (egoísta y recíproco) se hace más pequeña a medida que el sistema tiende a la configuración óptima. De esta forma, en las fases iniciales del aprendizaje se observa que el sistema tiende rápidamente a sustituir los comportamientos no adecuados por el comportamiento óptimo para cada ambiente. Sin embargo, no se aprecian diferencias entre ambas escalas, en lo referente a cual es la configuración óptima para cada ambiente, ni diferencias en la estabilidad de esa combinación óptima. Esto confirma la hipótesis de que el tamaño del sistema no afecta al mecanismo de aprendizaje propuesto ni a la estabilidad de la cooperación altruista mediante la estrategia de reciprocidad. En todos los casos se obtienen rendimientos cercanos al óptimo y estables.

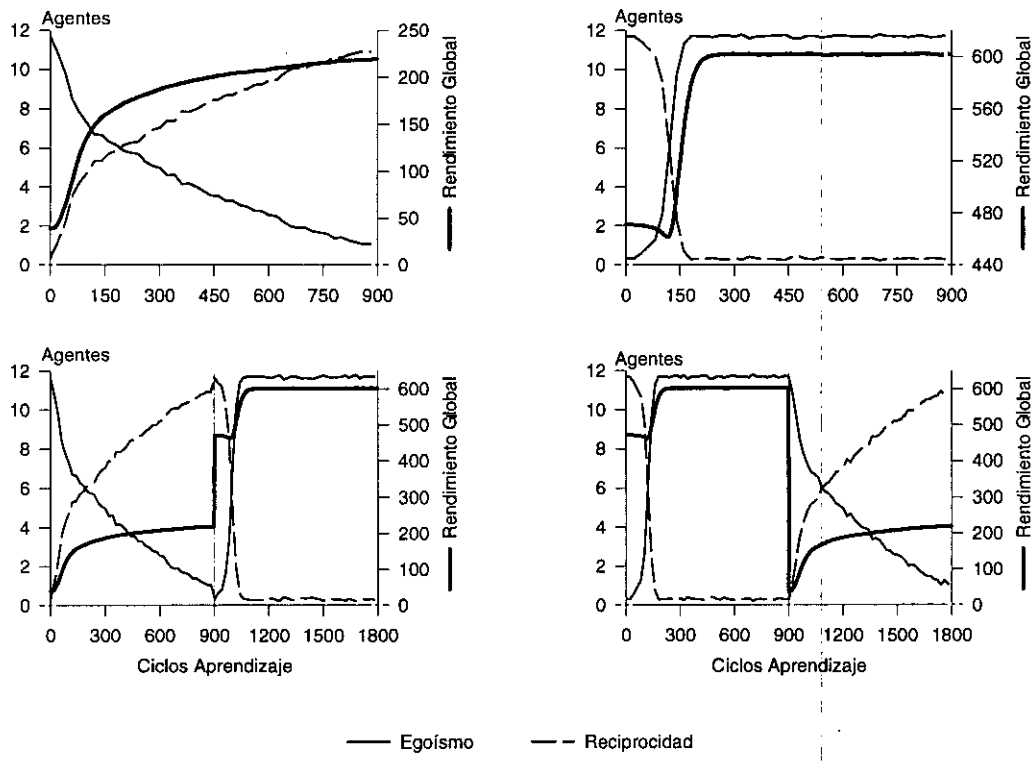


Figura 5.6. Resultados medios de 100 repeticiones de los experimentos de aprendizaje en el ambiente A (izquierda arriba), ambiente B (derecha arriba), ambiente C(izquierda abajo) y en el ambiente D (derecha abajo). Se muestra el rendimiento global del sistema y el número de agentes exhibiendo cada una de las dos alternativas de comportamiento cooperativo (egoísta, recíproca). En los ambientes C y D, el cambio de distribución de objetos ocurre en el paso 900.

Tabla 5.9. Valores medios y desviaciones típicas sobre 100 repeticiones del rendimiento global (R_g) al inicio y final de la fase de aprendizaje y número de agentes mostrando cada alternativa de comportamiento al final del aprendizaje.

	Ambiente A			
	Rendimiento Global		Nº medio de agentes por comportamiento	
	$\overline{x}$	s	Egoísta	Recíproco
Inicial	38,96	12,47	1,04	10,46
Final	220,70	10,90		
	Z = -8,68.	p<0,001	IC(99%) $p_{recíprocos}$ = 0,913 ± 0,020	
	Ambiente B			
	Rendimiento Global		Nº medio de agentes por comportamiento	
	$\overline{x}$	s	Egoísta	Recíproco
Inicial	470,76	9,07	11,67	0,33
Final	603,40	21,72		
	Z = -8,67.	p<0,01	IC(99%) $p_{recíprocos}$ = 0,275 ± 0,012	
	Ambiente C			
	Rendimiento Global		Nº medio de agentes por comportamiento	
	$\overline{x}$	s	Egoísta	Recíproco
Cambio	470,62	9,04	11,75	0,25
Final	604,25	17,70		
	Z = -8,68.	p<0,01	IC(99%) $p_{recíprocos}$ = 0,021 ± 0,01	
	Ambiente D			
	Rendimiento Global		Nº medio de agentes por comportamiento	
	$\overline{x}$	s	Egoísta	Recíproco
Cambio	36,93	12,19	0,93	11,07
Final	218,99	10,78		
	Z = -8,68.	p<0,01	IC(99%) $p_{recíprocos}$ = 0,915 ± 0,021	

5.6 Efecto de la competición por los recursos

La competición por los recursos es una situación frecuente en los seres vivos y se manifiesta tanto a nivel intraespecífico como interespecífico. En las sociedades animales se compite por el espacio, por el alimento, por la reproducción, etc. La competencia es un factor más que se integra en el conjunto de las presiones selectivas que sufren los seres vivos. En el caso de los comportamientos sociales, la escasez de recursos modifica la utilidad de distintas estrategias de cooperación. Las diferencias que existen entre los beneficios de una alternativa no social egoísta y los costes de una alternativa altruista se hacen más acentuadas (Zamora, Millán y Murciano, 1997).

El planteamiento experimental utilizado hasta el momento dispone un número inagotable de objetos en el ambiente de trabajo. Por esta razón, el sistema multiagente desarrolla su tarea de forma que no existe competición por los recursos entre los agentes. En los experimentos que se exponen en este apartado, se ha limitado el número de objetos, de forma que aparece la competición entre los agentes por los objetos. Esto provoca que la cooperación tenga un coste adicional. No sólo se paga por el tiempo que un agente pierde por estar detenido señalizando, sino que además, los objetos que recoja otro agente fruto de la ayuda recibida desaparecen del ambiente pudiendo llegar a su extinción antes de que termine el periodo de recogida.

5.6.1 Pruebas sin aprendizaje.

Los experimentos de esta serie utilizan los ambientes A, B, C y D (Tabla 5.1) con la particularidad de que los objetos se han limitado a 45 en el ambiente A (que corresponde con la media de los objetos recogidos por todas las combinaciones posibles de comportamientos en ese ambiente) y 100 objetos en el ambiente B. Se realizan las mismas pruebas que en los apartados 5.1, 5.2 y 5.3, para probar, en las mismas circunstancias, el efecto de la competencia sobre la estabilidad del altruismo, sobre su estabilización mediante la reciprocidad y sobre la flexibilidad del algoritmo de aprendizaje.

Tabla 5.10. Resultados de 100 simulaciones sin aprendizaje en los ambientes A y B con efecto de competencia. Se presentan las medias y desviaciones típicas de los rendimientos individuales (R_a, R_e y R_r) y global (R_g) para cada combinación posible de los tres comportamientos en el equipo de 4 agentes. Con líneas de puntos se representa la ausencia de un comportamiento en la combinación.

Ambiente A

	R_g		R_a		R_e		R_r	
	$\overline{x}$	s	$\overline{x}$	s	$\overline{x}$	s	$\overline{x}$	s
AAAA	44,90	0,42	11,23	0,10	----	----	----	----
AAAR	44,66	1,41	11,21	0,50	----	----	11,04	1,09
AAAE	44,44	2,22	9,66	0,75	15,46	1,92	----	----
AARR	44,78	1,04	11,25	0,76	----	----	11,14	0,75
AARE	43,06	6,50	10,14	1,76	12,44	2,00	10,34	2,09
AAEE	44,00	2,36	7,12	0,85	14,88	1,18	----	----
ARRR	44,40	3,00	11,10	1,30	----	----	11,10	0,82
ARRE	42,96	5,12	10,74	1,80	10,38	2,26	10,92	1,39
AREE	40,10	6,36	8,44	1,77	11,62	2,15	8,42	1,68
AEEE	32,02	11,44	0,08	0,27	10,65	3,81	----	----
RRRR	44,82	0,69	----	----	----	----	11,21	0,17
RRRE	42,50	3,99	----	----	7,60	1,93	11,63	1,16
RREE	30,84	5,49	----	----	6,43	1,28	8,99	1,98
REEE	13,56	2,94	----	----	3,81	0,84	2,12	1,26
EEEE	11,28	2,67	----	----	2,82	0,67	----	----

Ambiente B

	R_g		R_a		R_e		R_r	
	$\overline{x}$	s	$\overline{x}$	s	$\overline{x}$	s	$\overline{x}$	s
AAAA	35,78	6,43	8,94	1,60	----	----	----	----
AAAR	35,16	6,87	8,93	1,77	----	----	8,36	2,14
AAAE	60,86	4,35	2,76	1,54	52,56	2,88	----	----
AARR	35,40	6,90	9,16	1,80	----	----	8,54	1,97
AARE	62,78	4,05	3,98	1,68	50,10	2,97	4,72	1,41
AAEE	90,38	4,19	1,57	1,07	43,62	2,23	----	----
ARRR	34,92	6,44	9,22	1,84	----	----	8,56	1,67
ARRE	64,82	5,11	5,06	2,17	47,84	2,74	5,96	1,77
AREE	88,28	4,45	2,56	1,10	41,12	2,05	3,48	1,23
AEEE	98,28	1,87	0,06	0,23	32,74	0,62	----	----
RRRR	35,18	7,69	----	----	----	----	8,79	1,92
RRRE	65,12	4,25	----	----	45,84	2,98	6,42	1,42
RREE	86,16	4,10	----	----	38,93	2,07	4,15	0,71
REEE	97,52	2,98	----	----	29,02	1,43	10,46	4,56
EEEE	99,26	1,04	----	----	24,81	0,26	----	----

La Tabla 5.10 muestra los resultados de 100 simulaciones sin aprendizaje de todas las posibles combinaciones de las tres alternativas del comportamiento cooperativo. En ella se puede observar un comportamiento diferentes al que se presenta en ausencia de competencia (Tabla 5.2). El análisis de los rendimientos globales en ambos ambientes demuestra la existencia de diferencias significativas en función de la combinación de comportamientos presente (Prueba de Kruskal-Wallis $\chi^2 = 516{,}38$ $p<0{,}01$ para el ambiente A y $\chi^2 = 696{,}70$ $p<0{,}01$ para el ambiente B). Para el ambiente A, los efectos de la competencia se aprecian en dos aspectos. El primero de ellos es que los rendimientos globales para casi todas las combinaciones se satura alcanzando niveles próximos al óptimo posible (45 objetos). Esto es así puesto que en muchas combinaciones de comportamientos sobra tiempo de recogida. La combinación que más se aleja de este rendimiento global es la formada exclusivamente por agentes egoístas, cuyo rendimiento es menor del 30% del máximo. El segundo efecto de la competición por los recursos se tiene en el rendimiento individual de los agentes altruistas en su relación con el número de egoístas en la población. Ya vimos que, en ausencia de competencia, el coste del altruismo crecía a medida que disminuía el número de altruistas pues ese coste se distribuye entre menor número de agentes. Ahora, a ese incremento del coste hay que sumarle el efecto que tiene el hecho de la extinción de los objetos. Otra aspecto interesante relacionado con el rendimiento individual de las distintas estrategias cooperativas es que en ausencia de competición, el rendimiento individual de un agente cooperativo (recíproco o altruista) sólo depende del número de agentes cooperativos en la población (sin distinguir entre altruistas y recíprocos). Cuando existen efectos de competición, este rendimiento depende del número de cooperativos y de la presencia de altruistas entre ellos. En el siguiente ejemplo se clarifica este aspecto. Sea R_a(C) el rendimiento de un altruista en la combinación de comportamientos C, y R_r(C) el rendimiento individual de un recíproco en la misma combinación. En ausencia de competición se tiene que

$$R_a(\text{AREE})=R_r(\text{AREE})=R_a(\text{AAEE})=R_r(\text{RREE}).$$

El efecto de la competición convierte esa relación en la siguiente desigualdad

$$R_a(\text{AAEE})<R_a(\text{AREE})=R_r(\text{AREE})<R_r(\text{RREE}),$$

Esta desigualdad se corresponde con la siguiente afirmación: para un número fijo de agentes cooperativos (altruistas y recíprocos) el rendimiento individual de ambos tipos es el mismo y crece cuando disminuye el número de altruistas. Ello se debe a que los altruistas favorecen la recogida de objetos por los egoístas, aumentando la competencia contra los cooperativos.

En el ambiente B (objetos uniformemente distribuidos), los efectos de la competencia son menores y se relacionan exclusivamente con el rendimiento global que en casi todos los casos se acerca al óptimo habiendo mucha homogeneidad entre las distintas combinaciones. Al igual que en ausencia de competencia, el rendimiento individual de los agentes egoístas es siempre superior al del resto de alternativas de comportamiento.

5.6.2 Inestabilidad del altruismo en presencia de efectos de competición.

En este experimento, los agentes sólo pueden mostrar las alternativas de comportamiento cooperativo altruista y egoísta. Se inicializa sus variables de tendencia de comportamiento de forma que al principio del aprendizaje muestran todos el comportamiento altruista ($s_a(i,0)=k_s$ y $s_e(i,0)=0$, $\forall i$). El experimento se desarrolla en el ambiente A.

La Figura 5.7 muestra la evolución del rendimiento colectivo y el número de agentes mostrando cada una de las dos alternativas de comportamiento cooperativo a lo largo de los ciclos de aprendizaje.

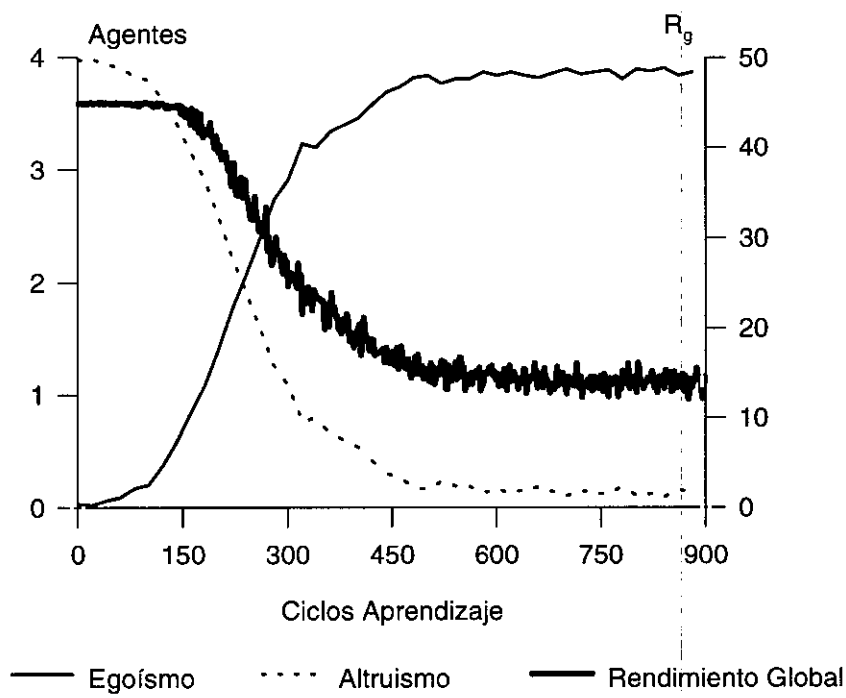


Figura 5.7. Resultados medios de 100 repeticiones del experimento de aprendizaje en el ambiente A en ausencia del comportamiento recíproco. Se muestra el rendimiento global del sistema y el número de agentes exhibiendo cada una de las dos alternativas de comportamiento.

Se observa en la Figura 5.7 que, por un lado el rendimiento global se mantiene durante un tiempo debido a que muchas de las combinaciones formadas por agentes altruistas y egoístas alcanzan rendimientos próximos al óptimo. Por otro lado, la invasión del egoísmo sucede antes que en el caso de ausencia de efectos de competencia. Esto es debido a que el coste de la cooperación se ve incrementado por el hecho de que una ayuda realizada a un individuo egoísta supone un objeto menos para recoger por los altruistas. Como consecuencia, la invasión egoísta en una población de altruistas sucede con mayor celeridad.

En la Tabla 5.11 se muestran los resultados numéricos del experimento. Las pruebas estadísticas (rangos con signo) muestran diferencias significativas en el rendimiento global en las fases inicial y final del aprendizaje, así como una diferente distribución de comportamientos al final del aprendizaje. Los resultados demuestran que el efecto de competencia provoca una mayor inestabilidad del altruismo puro al enfrentarse a la estrategia egoísta.

Tabla 5.11. Valores medios y desviaciones típicas sobre 100 repeticiones del rendimiento global (R_g) al inicio y final de la fase de aprendizaje y número de agentes mostrando cada alternativa de comportamiento al final del aprendizaje.

	Rendimiento Global		Nº medio de agentes por comportamiento	
	$\overline{x}$	s	Egoísta	Altruista
Inicial	44,87	0,48	3,83	0,17
Final	14,29	9,28		
	Z = -8,46.	p<0,01	IC(99%) $p_{altruista}$ = 0,032 ± 0,023	

5.6.3 Estabilización del altruismo mediante la estrategia TFT en presencia de efectos de competición.

En este experimento, los agentes pueden mostrar las tres alternativas de comportamiento cooperativo (egoísta, recíproco y altruista). Los agentes inicializan sus variables de tendencia de comportamiento de forma que al principio del aprendizaje muestran todos el comportamiento altruista ($s_a(i,0)=k_s$, $s_r(i,0)=0$ y $s_e(i,0)=0$, $\forall i$). El experimento se desarrolla en los ambientes A y B.

La Figura 5.8 muestra la evolución del rendimiento colectivo y el número de agentes mostrando cada una de las alternativas de comportamiento cooperativo según procede el aprendizaje. En la Tabla 5.12 se muestran los valores medios y desviaciones típicas, calculados sobre 100 repeticiones, del rendimiento global y del número de agentes en cada comportamiento en la fase final del aprendizaje. En el ambiente A, no se aprecian diferencias significativas en el rendimiento global entre la fase inicial del aprendizaje y el final del mismo, mientras que en el ambiente B, la prueba de rangos con signo demuestra un incremento significativo del rendimiento global. Asimismo, se presentan los intervalos de confianza la 99% para la proporción de recíprocos en ambos ambientes. De la observación de éstos se extrae la existencia de diferentes composiciones de comportamientos en el equipo en los ambientes A y B.

Tabla 5.12. Valores medios y desviaciones típicas sobre 100 repeticiones del rendimiento global (R_g) al inicio y final de la fase de aprendizaje y número de agentes mostrando cada alternativa de comportamiento al final del aprendizaje.

	Ambiente A				
	Rendimiento Global		N° medio de agentes por comportamiento		
	$\bar{x}$	s	Egoísta	Recíproco	Altruista
Inicial	44,73	0,80	0,09	3,59	0,32
Final	44,34	2,51			
	Z = -0,98.	P=0.323	IC(99%) $p_{cooperativos} = 0{,}977 \pm 0{,}019$		
	Ambiente B				
	Rendimiento Global		N° medio de agentes por comportamiento		
	$\bar{x}$	s	Egoísta	Recíproco	Altruista
Inicial	36,44	8,52	3,86	0,06	0,08
Final	98,91	1,23			
	Z = -8,68.	p<0,01	IC(99%) $p_{cooperativos} = 0{,}035 \pm 0{,}024$		

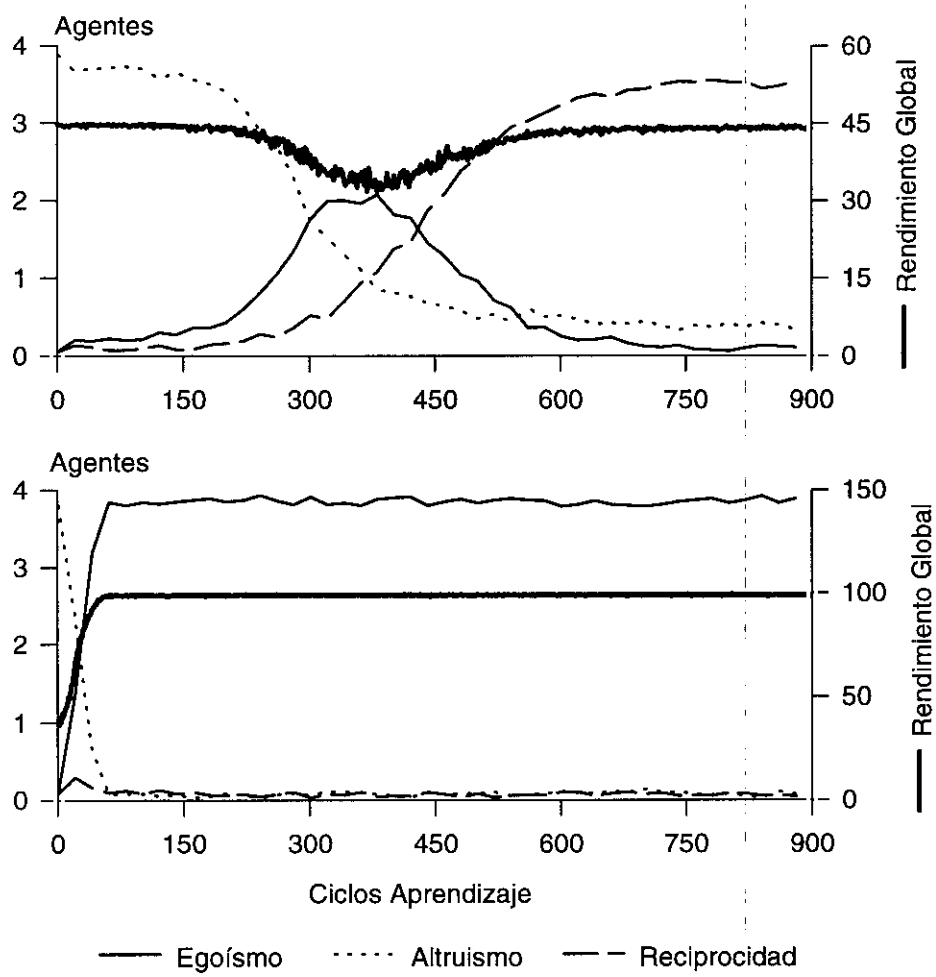


Figura 5.8. Resultados medios de 100 repeticiones de los experimentos de aprendizaje en el ambiente A (arriba) y en el ambiente B (abajo). Se muestra el rendimiento global del sistema y el número de agentes exhibiendo cada una de las alternativas de comportamiento cooperativo (egoísta, recíproca y altruista).

Pese a las diferencias en los rendimientos individuales y en el colectivo señaladas anteriormente, la inclusión de competición no se manifiesta con ningún efecto significativo en los resultados finales del proceso de aprendizaje. Sin embargo es importante reseñar que, durante el aprendizaje en el ambiente A, el sistema se recupera más rápidamente de la invasión por el comportamiento egoísta mediante la adopción de la estrategia recíproca. Esto es debido a que, como hemos visto (sección 5.5.1), la presencia de competencia provoca que el rendimiento individual de los agentes con comportamientos cooperativos

(ya sea recíprocos o altruistas) sea menor cuando el colectivo contiene algún agente altruista. Cuando desaparecen éstos por efecto de la invasión egoísta, los agentes seleccionan más frecuentemente la alternativa recíproca frente a la altruista recuperándose antes de la invasión. La traducción de esta invasión en el rendimiento global se ve amortiguada en comparación con los experimentos sin competencia, además de por el efecto de la suma de 100 repeticiones de invasiones en distintos tiempos, porque en casi todos los estados por los que pasa el sistema en su transición de altruismo puro a estabilización de la reciprocidad, se obtienen rendimientos cercanos al óptimo. Finalmente, tanto en ausencia como en presencia de competición por recursos, el sistema alcanza la configuración óptima y estable para ambos ambientes.

Tabla 5.13. Valores medios y desviaciones típicas sobre 100 repeticiones del rendimiento global (R_g) en el momento del cambio de ambiente y al final de la fase de aprendizaje y número de agentes mostrando cada alternativa de comportamiento al final del aprendizaje.

	Ambiente C				
	Rendimiento Global		N° medio de agentes por comportamiento		
	$\overline{x}$	s	Egoísta	Recíproco	Altruista
Cambio	37,23	10,89	3,85	0,09	0,06
Final	99,05	1,19			
	Z = -8,63. p<0,01		IC(99%) $p_{cooperativos}$ = 0,037 ± 0,024		
	Ambiente D				
	Rendimiento Global		N° medio de agentes por comportamiento		
	$\overline{x}$	s	Egoísta	Recíproco	Altruista
Cambio	12,23	6,07	0,07	3,78	0,15
Final	44,26	2,99			
	Z = -8,59. p<0,01		IC(99%) $p_{cooperativos}$ = 0,982 ± 0,017		

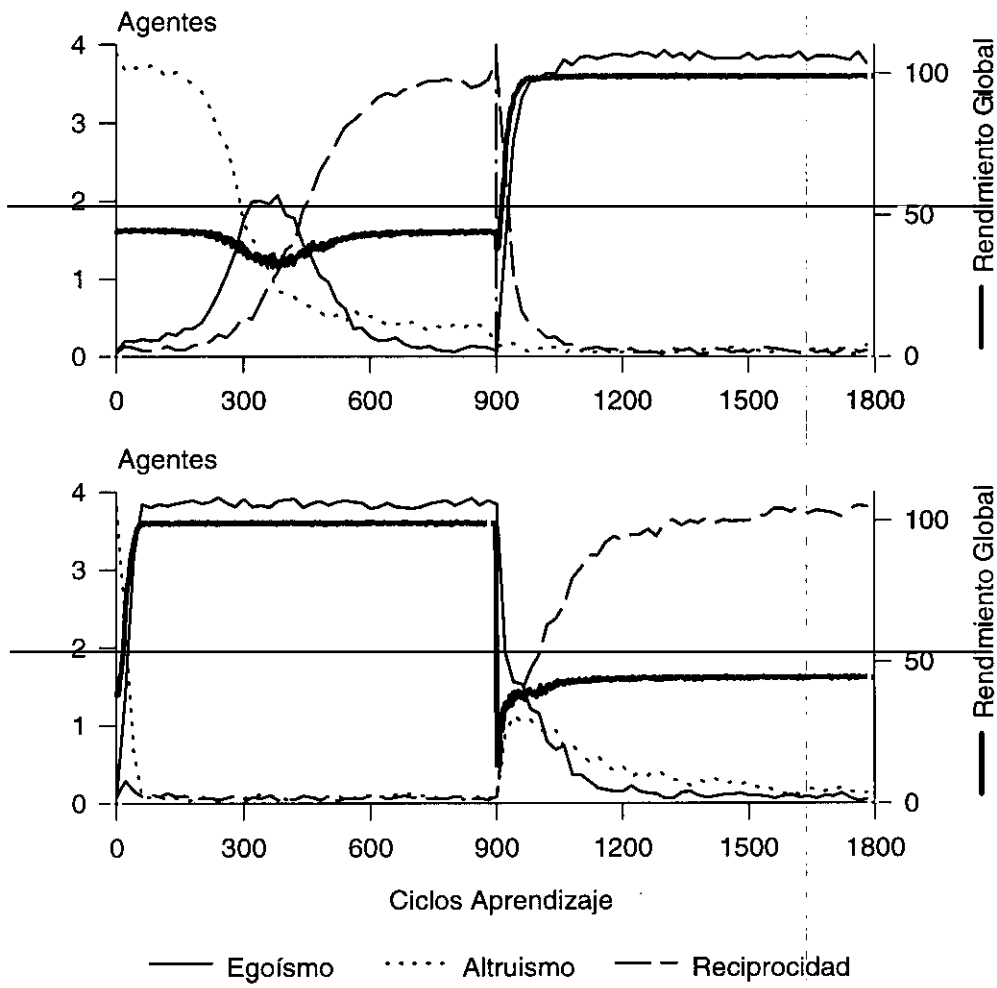


Figura 5.9. Resultados medios de 100 repeticiones de los experimentos de aprendizaje en el ambiente C (panel superior) y en el ambiente D (panel inferior). Se muestra el rendimiento global del sistema y el número de agentes exhibiendo cada una de las alternativas de comportamiento cooperativo (egoísta, recíproca y altruista). El cambio de ambiente se produce en el ciclo 900.

El comportamiento de los agentes frente al cambio de ambiente no se ve afectado por la inclusión de competencia por los recursos. Sólo se aprecian diferencias, como en el caso anterior, relacionadas con la velocidad del aprendizaje, que es superior en presencia de competencia. En la Figura 5.9 se muestra la evolución del rendimiento colectivo y el número de agentes mostrando cada una de las alternativas de comportamiento cooperativo a lo largo del aprendizaje en los ambientes C y D. En la Tabla 5.13 se muestran los valores medios y desviaciones típicas del rendimiento global en la fases inicial (en el momento del cambio de ambiente) y final del aprendizaje, así cmo el número medio de agentes en cada comportamiento al final del aprendizaje calculados sobre 100 repeticiones. Las pruebas estadísticas (rangos con signo) muestran diferencias significativas en el rendimiento global del sistema en el momento inmediatamente posterior al cambio de ambiente y el final del aprendizaje, así como una diferente distribución de comportamientos al final del aprendizaje.

5.7 Sumario del capítulo.

En este capítulo se han presentado los resultados de las simulaciones de un colectivo de agentes con la arquitectura AREA, desarrollando una tarea de recogida de objetos. Se han presentado los distintos ambientes de experimentación y se han caracterizado en términos de la eficacia de los agentes, individual y colectiva, para resolver la tarea propuesta. La estrategia de cooperación óptima depende del ambiente donde se desenvuelve el sistema. Se han presentado pruebas experimentales de la inestabilidad del altruismo puro cuando se enfrenta a la alternativa egoísta en un sistema con aprendizaje. Se ha demostrado que la inclusión en la arquitectura de los agentes del comportamiento de reciprocidad consigue estabilizar, cuando el ambiente lo requiere, la cooperación altruista en el colectivo. Esta estabilización se traduce en rendimientos cercanos al óptimo colectivo. Los resultados obtenidos en los distintos ambientes demuestran la adaptabilidad del sistema frente a cambios ambientales, que se traduce en que los agentes, por medio del aprendizaje, eligen y estabilizan en el colectivo la estrategia óptima para cada ambiente probado. Finalmente, se han realizado pruebas para estudiar la escalabilidad del sistema propuesto y su comportamiento ante la presencia de efectos de competición por los recursos. En ambos casos, el sistema alcanza rendimientos óptimos con combinaciones de comportamientos que son evolutivamente estables.

Capítulo 6

Análisis de los límites de la Reciprocidad

El origen y el fundamento de la presente tesis es el paralelismo existente, en objetivos y problemática, entre los sistemas naturales y artificiales. Concretamente, en el contexto de los comportamientos sociales como la cooperación, el problema de la estabilidad del altruismo es equivalente en ambas situaciones. Para su solución en los sistemas multiagente, se ha propuesto una estrategia inspirada en el comportamiento animal: la reciprocidad. Una vez analizada la evolución de la estrategia de reciprocidad en los sistemas multiagente con aprendizaje, se hace necesario el estudio de los requerimientos de estabilización de la misma para completar la revisión de este paralelismo.

La estabilidad evolutiva de la estrategia de reciprocidad se fundamenta en que el actor del comportamiento altruista compensa el coste de la cooperación, debido a que los individuos ayudados cooperan a su vez en un futuro encuentro, es decir "devuelven el favor". En este sentido, la reciprocidad se considera *temporalmente altruista*. Este planteamiento sufre, sin embargo, de tres aspectos críticos que condicionan el éxito evolutivo de dicha estrategia: i) reconocimiento del engaño, ii) viscosidad de la población y iii) ventaja de la cooperación. Este capítulo se dedica al análisis experimental de dichos condicionantes con el fin de determinar su influencia sobre la estabilidad del altruismo en un sistema multiagente. El principal objetivo es caracterizar el espacio de viabilidad de la cooperación recíproca en la arquitectura AREA.

Para los experimentos de este capítulo, se utiliza un colectivo de 12 agentes con la arquitectura AREA. Los agentes presentan las alternativas de comportamiento cooperativo denominadas recíproca y egoísta. El sistema multiagente trabaja en un ambiente de objetos no limitantes (inagotables) y agrupados en zonas de difícil acceso (ambiente A de los experimentos del capítulo 5). Dado que se quiere probar la estabilidad de la reciprocidad, se elige este ambiente pues en él se ha demostrado que la cooperación es ventajosa. En cada sección se especifican las modificaciones que se introducen a este planteamiento experimental general.

Mediante estos experimentos se pretende probar el efecto de los factores anteriormente expuestos, por un lado sobre el establecimiento, mediante el algoritmo de aprendizaje propuesto, de la estrategia de reciprocidad en el colectivo de agentes, y por otro lado, sobre la estabilidad de la estrategia recíproca una vez instaurada en el sistema multiagente.

6.1 Errores de reconocimiento.

La estrategia de reciprocidad denominada TFT consiste en cooperar con aquellos individuos que también cooperan, y evitar hacerlo con los que no lo hacen. Para poner en marcha esta estrategia, es necesario cooperar inicialmente con todos los individuos de la población y en posteriores interacciones repetir el comportamiento mostrado por el oponente. Para la viabilidad de esta estrategia se necesita de un reconocimiento preciso de los individuos no cooperantes para evitar ayudarles en el futuro. En las relaciones de cooperación entre seres vivos, el reconocimiento puede ser pasivo, como el caso de los mutualismos en los que huésped y hospedador viven juntos, o mecanismos activos basados en percepciones sensoriales (olfato, visión, etc.) La precisión de estos mecanismos determinará la oportunidad de que un individuo con un comportamiento egoísta pueda aprovecharse de los esfuerzos cooperativos del resto y extenderse en la población. En el capítulo anterior se demuestra que el comportamiento altruista puro no es una estrategia evolutivamente estable. La cooperación incondicional que muestran los individuos con esa alternativa de comportamiento es el caso extremo de falta de reconocimiento del engaño.

La arquitectura AREA de los agentes autónomos, incluye un módulo de comportamiento social denominado **`reconocimiento-de-señalizadores`** que desempeña el papel de reconocimiento del engaño. Este módulo basa su funcionamiento en la comunicación entre los agentes, que incluyen su número de identificación en los mensajes que envían. Cuando un agente recíproco recibe un mensaje de *petición de ayuda*, identifica al emisor de esta comunicación. Si es la primera vez que interacciona con él, se muestra cooperativo como consecuencia de la estrategia TFT. Pasado un tiempo (k_{tft}) sin haber recibido ayuda de este

agente (ya sea un *mensaje de relevo* o una *donación de ayuda*), lo cataloga como egoísta y evita cooperar con él en el futuro (ver la descripción de comportamientos en el capítulo 4 para más detalles). Si un agente recíproco no reconoce a un egoísta, en el futuro cooperará con él haciendo que éste incremente su rendimiento individual. Cuando el efecto del error en el reconocimiento supere cierto límite, el rendimiento de los agentes con comportamiento egoísta superará al de los agentes recíprocos, y como consecuencia, el altruismo recíproco dejará de ser estable.

6.1.1 Evaluación del efecto del error en el reconocimiento: experimentos sin aprendizaje.

Se han desarrollado experimentos en los que se introducen porcentajes variables de error en el reconocimiento. El porcentaje de error se interpreta como la probabilidad de que un agente egoísta sea clasificado erróneamente como cooperativo. Un nivel de error del 100% se traduce en que los agentes recíprocos catalogan como cooperativos a todos los agentes egoístas presentes en la población, comportándose igual que si presentasen la alternativa de comportamiento cooperativo altruista, pues cooperan incondicionalmente con todos los agentes al considerarlos cooperativos independientemente de su comportamiento real. Un 50% de error significa que la mitad de las veces que un agente debería catalogar a otro como egoísta no lo hace.

En la Tabla 6.1 se presentan los resultados del rendimiento global y rendimientos individuales para distinto número de agentes cooperativos y niveles de error en el reconocimiento del 20%, 50% y 80%. Se han elegido los porcentajes de error 20%, 50% y 80% como muestra representativa de bajo, medio y alto nivel de error, pues los resultados con niveles extremos (0% y 100%) ya han sido expuestos en el capítulo 5, y se corresponden, en el caso de ausencia de error (0%) con los resultados de la sección 5.4, y en el caso de un nivel de error en el reconocimiento del 100% se corresponden con los resultados de la sección 5.1 (con la diferencia en el número de agentes). En la tabla se presentan los valores medios y desviaciones típicas, obtenidos sobre 100 repeticiones, del rendimiento individual y colectivo para todas las posibles combinaciones de ambas alternativas del comportamiento cooperativo. Como se observa en la tabla, el rendimiento global aumenta en las dos dimensiones del experimento. A medida que se incrementa el nivel de error en el reconocimiento, el rendimiento global crece, al igual que cuando se incrementa el número de agentes cooperativos en el equipo. La Figura 6.1 muestra ese efecto bivariante sobre el rendimiento global.

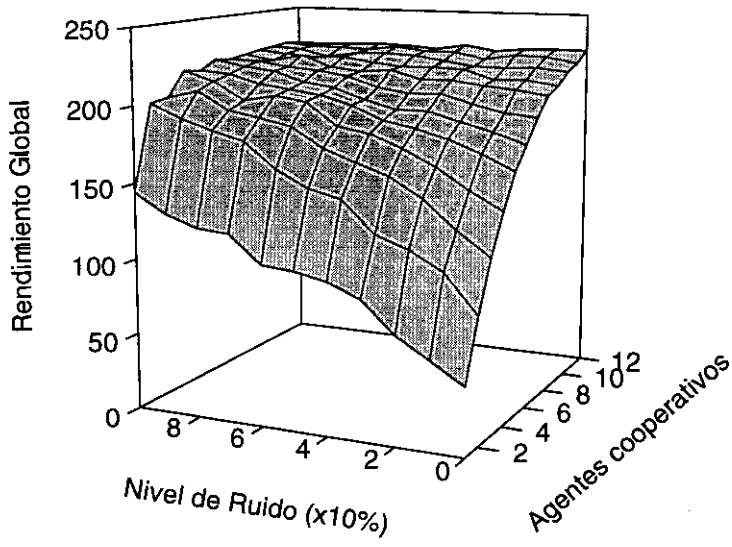


Figura 6.1. Rendimiento global medio, en 100 repeticiones, en función del nivel de error en el reconocimiento y del número de agentes cooperativos en el equipo.

Si atendemos por ejemplo a los resultados que se obtienen mediante un colectivo formado por igual proporción de agentes recíprocos y egoístas (Figura 6.2), se observa que el incremento en nivel de error en el reconocimiento provoca un aumento en el rendimiento global. Éste es, sin embargo, responsabilidad exclusiva de los agentes egoístas dado que a mayor error en el reconocimiento, mayor es también la probabilidad de que los agentes egoístas reciban ayuda de los recíprocos, con el consiguiente incremento en su rendimiento individual y su efecto directo sobre el rendimiento global. Por el contrario, los agentes recíprocos mantienen su rendimiento constante independientemente del nivel de error en el reconocimiento.

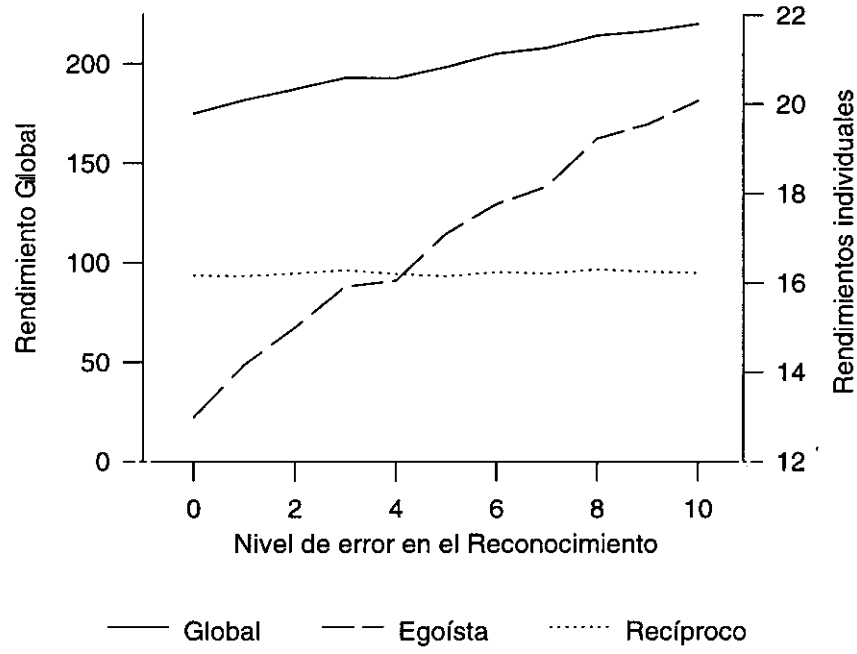


Figura 6.2. Valores medios sobre 100 repeticiones del rendimiento global y de los rendimientos individuales de cada uno de los comportamientos obtenidos por un colectivo de 12 agentes en el que la mitad muestra la alternativa de comportamiento egoísta.

En la Figura 6.3 se representan los rendimientos individuales de las dos alternativas de comportamiento según el nivel de error en el reconocimiento y el número de agentes cooperativos en el colectivo. En la figura se observa que el rendimiento individual de los agentes recíprocos (izquierda) no se ve afectado por el error en el reconocimiento[1]. Esto es así por la ausencia de competencia por los recursos. A pesar de que los agentes egoístas se aprovechan del error en el reconocimiento y reciben ayuda de los recíprocos, errar en el reconocimiento no se traduce en incrementos en el coste de la cooperación para los

[1] Puede apreciarse una pequeña variación en el rendimiento de un agente recíproco cuando éste es el único en el colectivo, entre las situaciones de ausencia y presencia de cualquier cantidad de error. Esto es debido al comportamiento peculiar que presenta un agente recíproco cuando se sabe aislado en la población. En ese caso, una vez que ha cooperado con todos los agentes de la población sin haber recibido ayuda, inhibe su comportamiento de `donación-de-ayuda`. Si existe error en el reconocimiento, el agente recíproco percibe erróneamente a otros cooperadores permaneciendo detenido durante todo el ciclo de recogida de objetos en espera de una señal de relevo.

recíprocos. En cambio, el rendimiento individual de los agentes egoístas (panel derecho) crece a medida que lo hace el error en el reconocimiento y cuando se incrementa el número de agentes cooperativos en el equipo de trabajo. Ambos factores incrementan la probabilidad de recibir ayuda por parte de los agentes cooperativos.

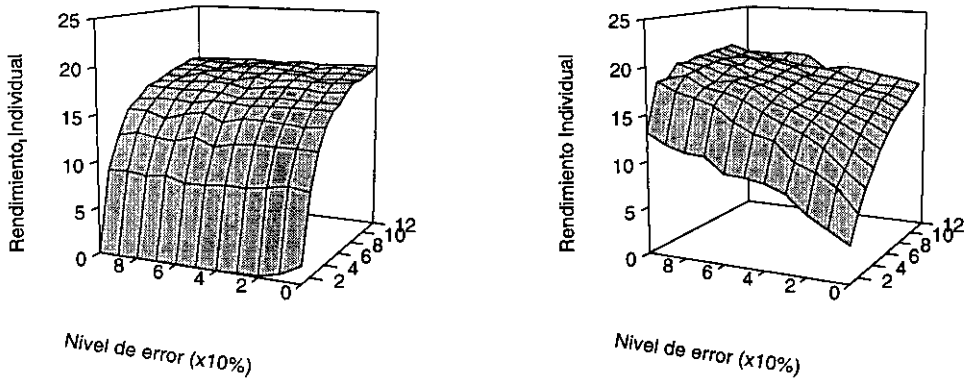


Figura 6.3. Rendimiento medio de un agente recíproco (izquierda) y un egoísta (derecha) en 100 repeticiones, en función del nivel de error en el reconocimiento y del número de agentes cooperativos en el colectivo.

6.1.2 Influencia del error en el reconocimiento sobre la estabilidad de la reciprocidad: Experimentos con aprendizaje.

El análisis de los rendimientos individuales y global según los distintos niveles de error en el reconocimiento, nos permite estudiar la estabilidad evolutiva de la reciprocidad y analizar el proceso de aprendizaje en función de los mismos.

Sea $R_i(m-n,n)$ el rendimiento individual de un agente con comportamiento $i \in$ {egoísta, recíproco}, en una configuración con m agentes de los que m-n son egoístas y n son recíprocos. El proceso de aprendizaje evalúa los rendimientos individuales de las dos

estrategias y según esta evaluación selecciona la estrategia que obtiene mayor rendimiento individual. Formalmente, un agente recíproco continuará mostrando dicho comportamiento si $R_r(m-n,n)>R_e(m-n+1,n-1)$.

La estabilidad evolutiva de una estrategia (recíproca en este caso) depende de su rendimiento individual pero en comparación con el rendimiento de la estrategia hipotéticamente invasora (egoísta). Una vez instaurada esta estrategia en el conjunto de los agentes, es decir cuando $n=m$, será resistente a invasiones de la estrategia egoísta si cumple una de las siguientes desigualdades:

$R_r(0,m)>R_e(1,m-1)$

o bien

$R_r(0,m)=R_e(1,m-1)$ y para una proporción pequeña p

$R_r(pm,(1-p)m)>R_e(pm+1,(1-p)m-1)$

En la Figura 6.4 se presentan los resultados comparativos de los rendimientos individuales de las estrategias recíproca y egoísta. La gráfica de la izquierda revela, de forma cualitativa, las zonas en las que el rendimiento de la estrategia egoísta es superior al correspondiente de la estrategia recíproca (zonas sombreadas). Es decir, se representa, para cada nivel de error y para cada posible combinación de comportamientos en el colectivo, el signo de la diferencia

$$R_e(m-n+1,n-1) - R_r(m-n,n).$$

La zonas sombreadas determinan los niveles de error para los que un agente recíproco de un equipo con n agentes recíprocos mejoraría su rendimiento individual si cambiara a comportarse de forma egoísta.

En la Figura 6.4 se observa que a partir de niveles de error en el reconocimiento superiores al 50%, la reciprocidad puede ser invadida por el egoísmo pues la rentabilidad de este último es superior. En el gráfico de la derecha se representa de forma cuantitativa las diferencias entre ambas alternativas de comportamiento cooperativo.

Tabla 6.1. Resultados de 100 simulaciones sin aprendizaje con niveles de error en el reconocimiento del 20%, 50% y 80%. Se presentan las medias y desviaciones típicas de los rendimientos individuales (R_e y R_r que corresponden a egoísta y recíproco respectivamente) y del global (R_g) para las combinaciones posible de ambos comportamientos en el equipo de 12 agentes.

	20% de error						**50% de error**						**80% de error**					
	R_g		R_r		R_e		R_g		R_r		R_e		R_g		R_r		R_e	
	$\overline{x}$	s	$\overline{x}$	s	$\overline{x}$	s	$\overline{x}$	s	$\overline{x}$	s	$\overline{x}$	s	$\overline{x}$	s	$\overline{x}$	s	$\overline{x}$	s
0	30,96	6,24	----	----	2,55	0,32	31,16	5,41	----	----	2,60	0,45	31,32	5,12	----	----	2,64	0,53
1	70,02	30,14	0,14	0,61	6,35	2,75	103,12	39,12	0,00	0,00	9,37	3,56	124,48	49,56	0,00	0,00	11,32	4,51
2	118,98	22,22	8,61	2,06	10,18	1,95	151,92	41,03	8,15	2,60	13,56	3,64	184,06	32,55	8,41	2,11	16,72	2,89
3	147,44	20,86	12,38	1,89	12,26	1,93	174,30	22,70	11,78	2,08	15,44	2,03	194,88	28,88	12,16	2,25	17,60	2,55
4	163,90	20,89	14,41	2,03	13,28	1,83	185,68	18,78	14,03	1,59	16,20	1,80	199,10	23,24	13,91	1,96	17,93	2,05
5	178,74	13,05	15,74	1,20	14,29	1,28	197,18	17,50	15,75	1,62	16,92	1,55	207,98	15,65	15,68	1,37	18,51	1,55
6	187,30	16,32	16,21	1,28	15,01	1,78	198,52	14,91	15,99	1,41	17,10	1,37	214,48	12,02	16,52	1,25	19,23	1,27
7	196,86	10,72	17,10	1,02	15,43	1,28	204,74	14,76	16,74	1,46	17,51	1,46	215,62	14,29	17,15	1,37	19,11	1,39
8	206,66	14,85	17,68	1,24	16,30	1,81	214,32	10,55	17,83	0,96	17,92	1,36	221,82	11,09	17,85	0,99	19,76	1,28
9	208,70	11,98	17,66	1,03	16,59	1,52	214,80	11,20	17,94	0,93	17,77	1,79	218,78	13,29	17,86	1,21	19,35	1,53
10	214,76	13,99	18,10	1,18	16,86	2,17	220,90	10,26	18,41	0,92	18,42	1,91	217,64	8,92	17,89	0,79	19,35	1,66
11	221,82	9,41	18,55	0,80	17,74	2,73	222,22	9,56	18,56	0,83	18,08	2,69	222,42	11,90	18,44	0,98	19,60	2,60
12	224,32	9,16	18,69	0,76	----	----	222,16	12,83	18,51	1,07	----	----	223,38	10,37	18,62	0,86	----	----

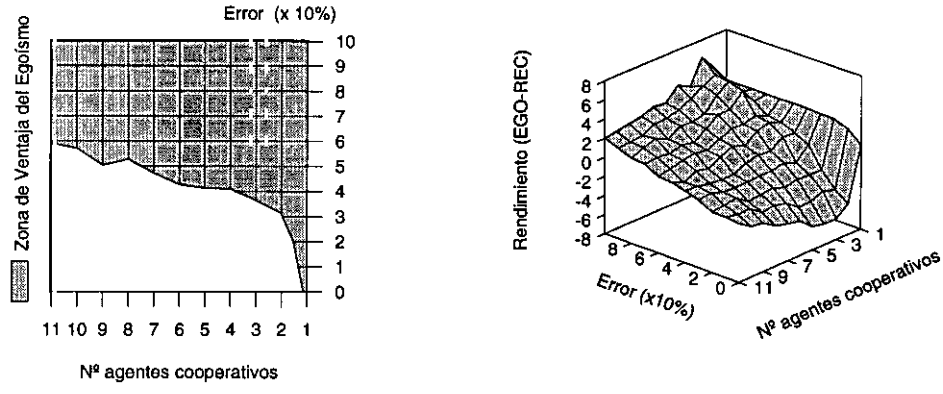


Figura 6.4. Análisis comparado del rendimiento individual de un agente recíproco y uno egoísta. En el panel de la izquierda se presenta, de forma cualitativa, el área donde el rendimiento comparado (Egoísta – Recíproco) es favorable al egoísmo. En el panel de la derecha se representa cuantitativamente este rendimiento comparado.

Los experimentos con aprendizaje se realizan en el mismo ambiente de objetos agrupados. Se han probado dos niveles de error en el reconocimiento (20% y 80%) en un sistema multiagente con 12 agentes. Estos pueden mostrar dos alternativas de comportamiento cooperativo: egoísta y recíproca.

La inicialización de las tendencias a cada una de ellas se realiza de dos maneras distintas. La primera es una inicialización *inespecífica*, en la que el agente decide su comportamiento inicial de forma aleatoria. La segunda inicialización, que denominaremos *adversa*, determina que inicialmente todos los agentes muestran la misma alternativa de comportamiento. Esta alternativa será distinta en cada caso, eligiendo siempre la alternativa contraria a la alcanzada al final del aprendizaje mediante inicialización inespecífica, en las mismas condiciones de nivel de error en el reconocimiento. Las variables de tendencia inicial para $i \in \{1,2,3,\ldots,12\}$ son, $s_e(i,0)=s_r(i,0)=k_s/2$ para el caso de inicialización *inespecífica*, y $s_j(i,0)=0$, $s_k(i,0)=k_s$, para la inicialización *adversa*, siendo j la alternativa de comportamiento estabilizada en las mismas condiciones con la inicialización inespecífica.

6.1.2.1 Resultados del aprendizaje con inicialización inespecífica.

Los resultados de los experimentos de aprendizaje con inicialización inespecífica para los dos niveles de error en el reconocimiento se presentan en la Figura 6.5 y en la Tabla 6.2 y Tabla 6.3.

En la Figura 6.5 se comienza el aprendizaje desde una situación en la que existe una proporción igual de ambas alternativas de comportamiento (egoísta y recíproco) en el colectivo de agentes. Con nivel de error 20% (panel superior), se observa inicialmente un rendimiento colectivo por debajo del óptimo que se obtuvo en las pruebas sin aprendizaje con un equipo de agentes trabajando en las mismas condiciones de error en el reconocimiento (Tabla 6.1). A medida que progresa el aprendizaje, los agentes comienzan a seleccionar la alternativa recíproca desplazando a la egoísta de la población. Esta sustitución se traduce en un incremento progresivo del rendimiento global. En la fase final del aprendizaje, el rendimiento global es cercano al óptimo y los agentes estabilizan su comportamiento mostrando todos la alternativa recíproca. En la Tabla 6.2 se muestran los resultados del rendimiento global en las fases inicial y final del aprendizaje y el número medio de agentes en cada comportamiento al final del aprendizaje. Una prueba de rangos con signo determina que las diferencias observadas en el rendimiento global son estadísticamente significativas. Se muestra también la estimación por intervalos de la proporción de agentes recíprocos al final del aprendizaje.

Tabla 6.2. Resultados obtenidos en 100 repeticiones con un nivel de error en el reconocimiento del 20% e inicialización inespecífica. Se muestran lo valores medios y desviaciones típicas del rendimiento global (R_g) al principio y al final de la fase de aprendizaje y el promedio de agentes en cada comportamiento al final del aprendizaje.

	Rendimiento Global		N° medio de agentes por comportamiento	
	$\overline{x}$	s	Egoísta	Recíproco
Inicial	188,39	22,48	0,30	11,70
Final	221,78	9,78		
	Z = -8,61	p<0,01	I.C(99%) $p_{recíproco}$ = 0,975±0,011	

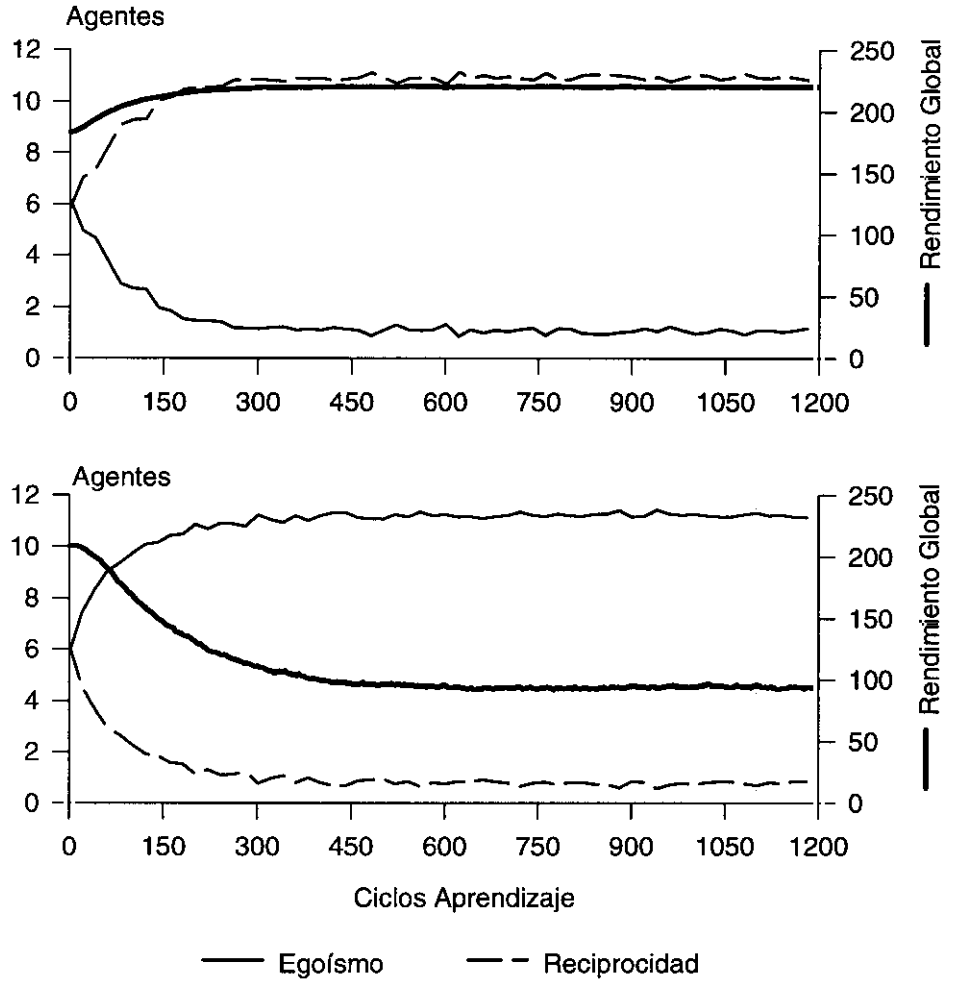


Figura 6.5. Inicialización inespecífica. Resultados de los experimentos con aprendizaje en función del error en el reconocimiento (20% en el panel superior y 80% en el panel inferior). Se muestra los valores medios sobre 100 repeticiones, del rendimiento global y del número de agentes mostrando cada una de las dos alternativas de comportamiento (egoísta y recíproca).

El resultado cambia cuando el error en el reconocimiento alcanza el 80% (Figura 6.5 inferior). El rendimiento global en estas circunstancias es cercano al óptimo obtenido en las pruebas sin aprendizaje. Sin embargo, a medida que progresa el aprendizaje, los agentes seleccionan la alternativa egoísta pues obtienen un alto rendimiento individual al recibir la cooperación de agentes recíprocos sin pagar ningún precio por ello. En cambio, el rendimiento individual de los agentes recíprocos es bajo pues inicialmente se distribuyen los costes de la cooperación entre un promedio de 6 agentes y a medida que éstos abandonan la cooperación al seleccionar la alternativa egoísta, el coste se incrementa. Esta diferencia entre rendimientos individuales provoca una rápida invasión del egoísmo que cuando se completa se traduce en una pérdida del rendimiento global situando éste muy por debajo del óptimo. Los resultados de la prueba de rangos con signos en la Tabla 6.3 muestran que la pérdida de rendimiento global entre las fases inicial y final del aprendizaje es altamente significativa. El intervalo de confianza para la proporción de agentes recíprocos en el colectivo al final del aprendizaje muestra que los agentes presentan una gran afinidad hacia la alternativa egoísta. El análisis de la desviación típica del rendimiento global al final del aprendizaje ilustra el importante efecto que sobre esta medida tiene el hecho de que alguno de los agentes pruebe la alternativa recíproca por efecto de un error en la selección de comportamiento. Este agente en su ciclo exploratorio cooperará con los agentes egoístas (un 80% de las veces) con el consiguiente incremento en el rendimiento individual de los mismos y por tanto en el rendimiento global. Esto provoca una mayor variabilidad en el resultado final de rendimiento global.

A la vista de las estimaciones realizadas de las proporciones de agentes recíprocos en las dos situaciones de error en el reconocimiento se extrae que la presencia de error en el reconocimiento del 80% provoca que los agentes seleccionen de forma muy significativa la alternativa de comportamiento egoísta.

Tabla 6.3. Resultados obtenidos en 100 repeticiones con un nivel de error en el reconocimiento del 80% e inicialización inespecífica. Se muestran lo valores medios y desviaciones típicas del rendimiento global (R_g) al principio y al final de la fase de aprendizaje y el promedio de agentes en cada comportamiento al final del aprendizaje.

	Rendimiento Global		N° medio de agentes por comportamiento	
	$\bar{x}$	s	Egoísta	Recíproco
Inicial	211,64	17,44	10,99	1,01
Final	104,87	70,97		
	$Z = -8,24$	$p<0,01$	I.C(99%) $p_{recíproco} = 0,084 \pm 0,02$	

6.2.1.2 Resultados del aprendizaje con inicialización adversa.

La inicialización de las variables de tendencia a los comportamientos en esta sección se realiza en el sentido opuesto a la configuración obtenida en los experimentos con inicialización inespecífica. Así, en el caso de un error en el reconocimiento del 20%, la inicialización se realiza de forma que todos los agentes muestran inicialmente la alternativa de comportamiento egoísta. En cambio, en el caso de un error del 80%, el comportamiento inicial de los agentes es el recíproco. Esta inicialización pretende situar al sistema en un punto de partida lo más alejado posible de la situación de estabilidad alcanzada anteriormente. Así se comprueba la potencia del algoritmo de aprendizaje para revertir de esta situación adversa. En estos experimentos, además de estudiar la calidad del algoritmo de aprendizaje propuesto, se estudia la estabilidad de la solución alcanzada observando si ésta permanece de forma estable durante la fase final del aprendizaje.

En la Figura 6.6 se presentan los resultados del aprendizaje en presencia de niveles de error del 20% (panel superior) y 80% (panel inferior). Cuando el nivel de error en el reconocimiento es del 20%, el primer agente que decide mostrar el comportamiento cooperativo recíproco obtiene un rendimiento individual mucho menor del que presentaba cuando se comportaba egoístamente (Tabla 6.1). Se presenta por tanto un problema de viabilidad inicial de la cooperación. El algoritmo de aprendizaje propuesto, mediante su factor de exploración variable, permite superar esta dificultad inicial de la instauración de la estrategia recíproca. En el momento en que aparecen simultáneamente dos agentes cooperativos en el sistema, éstos mejoran en gran medida su rendimiento individual, decidiendo cambiar su tendencia de comportamiento y comenzar a mostrar con más frecuencia la alternativa recíproca. Este paso inicial de exploración, cataliza el proceso de sustitución gradual del comportamiento egoísta por la nueva alternativa más rentable, la recíproca. El efecto de esta sustitución de comportamientos en la configuración del equipo de trabajo, se traduce en un incremento gradual del rendimiento global del sistema llegándose, al final del aprendizaje, a una situación estable en comportamientos y cercana al óptimo en rendimiento global. La Tabla 6.4 muestra los resultados medios sobre 100 repeticiones del rendimiento global en las fases inicial y final del aprendizaje y el promedio de agentes en cada alternativa de comportamientos al final del aprendizaje. En ambos casos, las diferencias observadas son estadísticamente significativas.

Cuando el error en el reconocimiento es del 80% (Figura 6.6-inferior), los agentes parten de una situación de reciprocidad. Los rendimientos individuales de los agentes egoístas en esa situación son muy superiores a los obtenidos por los agentes recíprocos, de forma que la mejor estrategia que puede elegir un agente es la egoísta. Por esta razón, los agentes seleccionan esta alternativa invadiendo paulatinamente la población. El rendimiento colectivo se mantiene durante algún tiempo en niveles cercanos al óptimo, dado que con un 80% de error en el reconocimiento, la mayoría de los agentes egoístas reciben cooperación de los agentes recíprocos que se mantienen en la población. Cuando la invasión del

egoísmo es completa, el rendimiento global decrece notablemente situándose por debajo del 50% del obtenido en la fase inicial del aprendizaje. Es importante notar la presencia de gran variabilidad en el rendimiento global al final del aprendizaje pues, aunque el sistema se estabiliza en el comportamiento egoísta, el cambio exploratorio de un agente hacia el comportamiento recíproco se traduce en un incremento importante en el rendimiento individual de los agentes egoístas. La Tabla 6.5 muestra los resultados medios sobre 100 repeticiones del rendimiento global en las fases inicial y final del aprendizaje y el promedio de agentes en cada alternativa de comportamiento al final del aprendizaje. La prueba de rangos con signo concluye que las diferencias observadas en el rendimiento global del sistema son estadísticamente significativas. El intervalo de confianza al 99% determina que los agentes seleccionan de forma muy significativa la alternativa de comportamiento recíproca.

Tabla 6.4. Resultados obtenidos en 100 repeticiones con un nivel de error en el reconocimiento del 20% e inicialización *adversa* en egoísmo. Se muestran lo valores medios y desviaciones típicas del rendimiento global (R_g) al principio y al final de la fase de aprendizaje y el promedio de agentes en cada comportamiento al final del aprendizaje.

	Rendimiento Global		N° medio de agentes por comportamiento	
	$\overline{x}$	s	Egoísta	Recíproco
Inicial	45,47	29,43	0,46	11,54
Final	221,93	10,78		
	Z = -8,68	p<0,01	I.C(99%) $p_{recíproco}$ = 0,962±0,014	

Tabla 6.5. Resultados obtenidos en 100 repeticiones con un nivel de error en el reconocimiento del 80% e inicialización *adversa* en reciprocidad. Se muestran lo valores medios y desviaciones típicas del rendimiento global (R_g) al principio y al final de la fase de aprendizaje y el promedio de agentes en cada comportamiento al final del aprendizaje.

	Rendimiento Global		N° medio de agentes por comportamiento	
	$\overline{x}$	s	Egoísta	Recíproco
Inicial	222,52	10,67	11,19	0,81
Final	95,57	68,61		
	Z = -8,65	p<0,01	I.C(99%) $p_{recíproco}$ = 0,067±0,018	

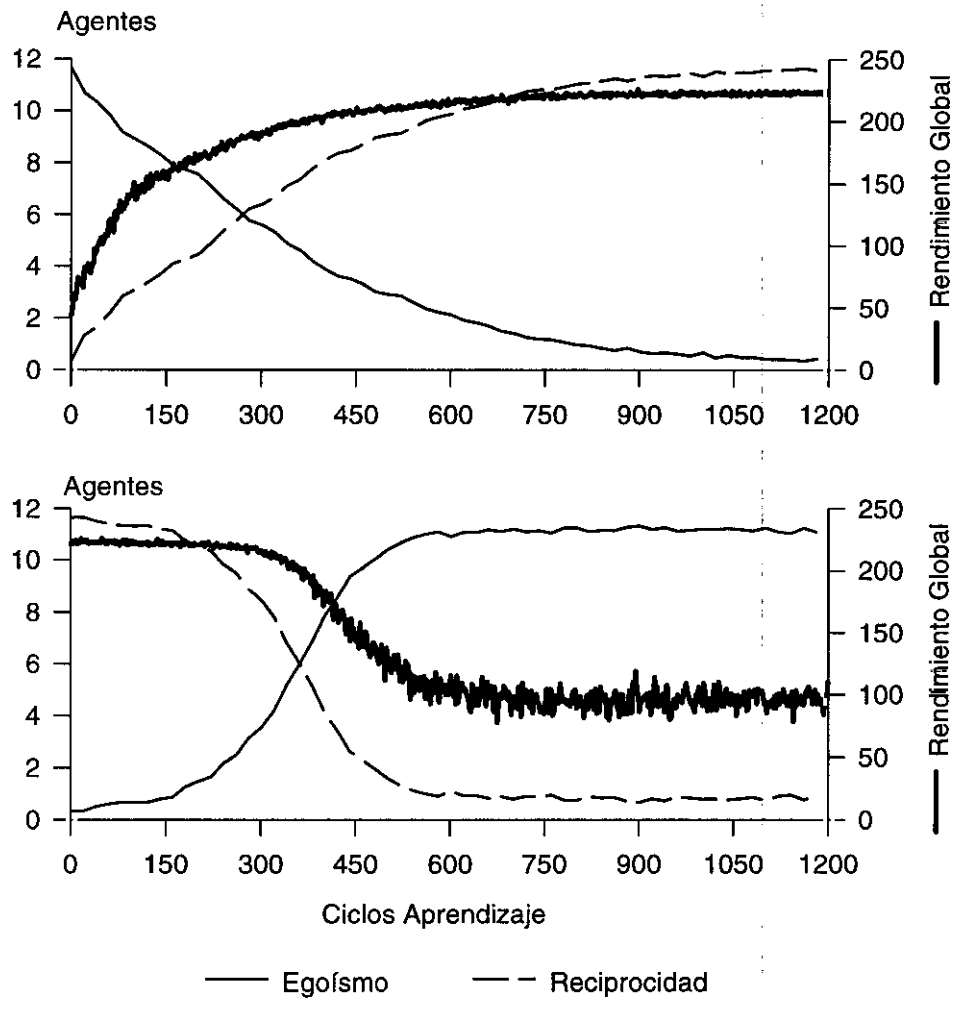


Figura 6.6. Inicialización egoísta. Resultados de los experimentos con aprendizaje en función del error en el reconocimiento (20% panel superior y 80% panel inferior). Se muestran los valores medios sobre 100 repeticiones, del rendimiento global y del número de agentes mostrando cada una de las dos alternativas de comportamiento(egoísta y recíproca).

6.1.3. Discusión

Tras el análisis de los resultados obtenidos en estos experimentos se constata la importancia del error en el reconocimiento en la estabilización de la reciprocidad. En presencia de un porcentaje de error reducido (20% o menor como en el caso de la sección 5.4 que se corresponde con un error del 0%), el sistema es capaz de instaurar la cooperación mediante reciprocidad sea cual sea el punto de partida de las tendencias de comportamiento de los agentes (Tabla 6.2 y Tabla 6.4). Una vez instaurada, permanece de forma estable en el colectivo resistiendo la invasión de la estratega egoísta, y produciendo un rendimiento global cercano al óptimo. Por el contrario, si el nivel de error es del 80% (o del 100% como se podría interpretar en el caso del comportamiento altruista puro de la sección 5.1), la reciprocidad deja de ser estable y el egoísmo invade la población provocando que el rendimiento global del sistema esté muy por debajo del óptimo. Esta invasión se produce tanto en los experimentos de inicialización inespecífica como en los experimentos en los que el sistema se inicializa en la reciprocidad (Tabla 6.3 y Tabla 6.5).

Es importante notar que la variabilidad en el rendimiento global de las distintas fases del aprendizaje es mayor cuando el sistema presenta mayoritariamente el comportamiento egoísta, ya sea por inicialización en esta situación, o bien porque el sistema la alcance mediante el aprendizaje. El cambio de un agente egoísta a recíproco en esa situación provoca que el resto de individuos mejoren momentáneamente su rendimiento individual provocando un incremento en el rendimiento global.

En la naturaleza se encuentran diversos mecanismos para el reconocimiento del engaño. Dependiendo de la habilidad de una especie en realizar este reconocimiento, será necesario recurrir o no a otros mecanismos para evitar el abusos de las estrategias no cooperativas. Estos mecanismos van desde la simple interacción continuada con el mismo individuo por vivir junto a él, hasta mecanismos de territorialidad en los que los individuos de distintos grupos permanecen en territorios disjuntos evitando interaccionar entre ellos. En el caso que nos ocupa, el reconocimiento del engaño recae sobre uno de los comportamientos de la arquitectura AREA de los agentes, que permite el reconocimiento entre agentes basado en la comunicación.

A la vista de los resultados obtenidos, se puede asumir cierto nivel de error en la comunicación entre agentes sin que esto suponga un riesgo para la estabilidad de la reciprocidad. En el caso de la implementación en robots reales, el nivel de ruido en la comunicación es controlado mediante procesos de análisis de la consistencia de los mensajes. Esto es, los mensajes deben tener significado (ser alguno de los tipos diseñados en el sistema de comunicación) y debe ser reiterado, es decir, debe recibirse un número mínimo de veces el mensaje correcto para ser tenido en cuenta. Las evaluaciones realizadas con los robots reales indican que el error en la comunicación entre robots afecta a la

pérdida de mensajes más que a la mala interpretación de los mismos. En cualquier caso, este error en el reconocimiento nunca supera un porcentaje del 20%.

6.2 Viscosidad de la población

El éxito de la estrategia de reciprocidad TFT, se basa en la compensación del coste de la cooperación altruista por la recepción de ayuda en futuros encuentros. De la misma forma que la estabilidad del altruismo necesita de un mecanismo de reconocimiento del engaño para evitar abusos de la estrategia egoísta, es necesario que exista un número suficiente de encuentros entre los individuos que cooperan, para que sea posible amortizar el gasto invertido en la cooperación. Esto es, si dos agentes se encuentran en una única ocasión, la mejor estrategia posible, y a la par la única estable, es no cooperar. Pero si el número de encuentros aumenta y alcanza un valor lo suficientemente alto, la mejor estrategia posible es la reciprocidad. Este número de encuentros entre agentes cooperativos es lo que se conoce con el nombre de *viscosidad*. Hay varios factores que afectan al número de encuentros entre agentes. El más importante es el tiempo total de "vida" del agente, en nuestro caso, el tiempo que dura un ciclo de recogida de objetos (k_{t_limite}). Tiempos de recogida excesivamente cortos impiden que los agentes con comportamiento cooperativo recíproco compensen su gasto con futuras ayudas recibidas de reciprocidad. El número total de agentes en el colectivo también ejerce cierta influencia sobre la viscosidad. En colectivos de agentes muy numerosos se dan dos efectos que benefician a la estrategia egoísta acercando su eficacia a la que se obtiene mediante la reciprocidad. Por un lado, los agentes con comportamiento egoísta tienen mayores oportunidades de obtener beneficios de la primera cooperación de todos los estrategas TFT y, aunque de menor importancia, los agentes egoístas tienen mayor probabilidad de utilizar las señales de comunicación enviadas entre agentes recíprocos pues éstas son más numerosas.

Ahora bien, el efecto de abuso que ejercen los egoístas es menor cuando la duración del ciclo es mayor. En el límite, en duraciones de ciclo de recogida suficientemente grandes, el efecto de la primera cooperación de la estrategia TFT se diluye hasta hacerse despreciable recuperando los agentes recíprocos la ventaja atesorada por los agentes egoístas. En esta sección, se examina la influencia de la viscosidad sobre la estabilidad de la reciprocidad. El nivel de viscosidad se ha variado en los experimentos, definiendo distintos tiempos de simulación (k_{t_limite}). El objetivo de este estudio es determinar, de forma empírica, el efecto de distintas duraciones de ciclos de recogida sobre la estabilidad evolutiva de la reciprocidad.

6.2.1 Evaluación del efecto de la viscosidad: experimentos sin aprendizaje.

Se han desarrollado experimentos en los que se varía la duración de los ciclos de recogida, oscilando entre los 2000 y 10000 pasos. En la Tabla 6.6 se presentan los resultados del rendimiento global y rendimientos individuales sobre 100 repeticiones, para distinto número de agentes cooperativos y duraciones de la simulación de 2000 y 6000 pasos como representativos de ciclos de recogida cortos y largos respectivamente. En la Figura 6.7 se representa la evolución del rendimiento global en función del número de agentes cooperativos y de la duración de los ciclos de recogida.

El rendimiento global del sistema crece, como es natural, a medida que aumenta la duración de los ciclos de recogida. En cuanto a los rendimientos individuales de los agentes según su alternativa de comportamiento cooperativo (Figura 6.8), se observa que ambos tipos incrementan su rendimiento cuando aumenta la duración de los ciclos.

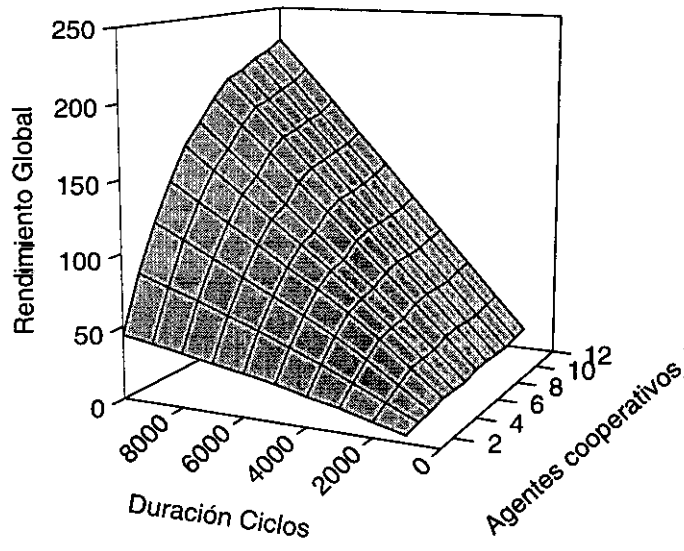


Figura 6.7. Rendimiento global medio, en 100 repeticiones, en función de la duración de los ciclos de recogida y del número de agentes cooperativos en el equipo.

Tabla 6.6. Resultados de 100 simulaciones sin aprendizaje con duración del ciclo de recogida de 2000 y 6000 pasos. Se presentan las medias y desviaciones típicas de los rendimientos individuales (R_e y R_r que corresponden a egoísta y recíproco respectivamente) y del global (R_g) para las combinaciones posible de ambos comportamientos en el equipo de 12 agentes.

	2000 Pasos						**6000 Pasos**					
	R_g		**R_r**		**R_e**		**R_g**		**R_r**		**R_e**	
	$\overline{x}$	s	$\overline{x}$	s	$\overline{x}$	s	$\overline{x}$	s	$\overline{x}$	s	$\overline{x}$	s
0	5,80	2,46	----	----	0,48	0,20	19,72	4,43	----	----	1,64	0,37
1	9,14	5,93	0,00	0,00	0,83	0,54	28,44	8,35	0,60	0,99	2,53	0,72
2	15,74	11,50	0,71	0,81	1,43	1,01	50,56	10,21	4,13	2,00	4,23	0,73
3	21,46	12,32	1,39	1,11	1,92	1,03	72,72	11,74	6,84	1,75	5,80	0,85
4	23,82	13,39	1,61	1,14	2,18	1,14	86,54	11,65	7,89	1,68	6,87	0,77
5	25,34	14,06	1,83	1,27	2,31	1,16	95,58	15,42	8,48	1,81	7,60	1,05
6	31,86	10,60	2,47	0,97	2,84	0,85	108,06	10,59	9,62	1,01	8,39	0,94
7	31,40	10,18	2,42	0,93	2,89	0,80	113,80	11,31	9,79	1,15	9,06	0,91
8	32,14	10,63	2,56	0,93	2,93	0,91	116,92	13,30	9,89	1,23	9,45	1,32
9	34,48	10,21	2,77	0,90	3,19	0,95	123,42	11,06	10,43	1,03	9,86	1,21
10	35,72	7,88	2,93	0,67	3,23	0,89	125,94	8,43	10,62	0,75	9,89	1,33
11	38,24	8,10	3,15	0,70	3,58	0,91	129,82	9,62	10,83	0,87	10,72	1,74
12	36,46	7,92	3,04	0,66	----	----	129,52	8,67	10,79	0,72	----	----

Para el estudio de la estabilidad de la reciprocidad en función de las distintas duraciones de ciclos de recogida, es interesante conocer la diferencia relativa de rendimiento entre las estrategias que compiten. En la Figura 6.9 se presenta el análisis comparativo de los rendimientos individuales de ambas estrategias. En el panel de la izquierda, la zona sombreada marca el área de ventaja favorable al egoísmo (sección 6.1.2). Se observa que cuando la duración de ciclo es de 2000 pasos, la ventaja es favorable al egoísmo. En el caso particular de un colectivo con un sólo agente cooperativo, la comparación de rendimientos es favorable al comportamiento egoísta independientemente de la duración del ciclo. Esto es así debido a que este agente cooperativo aislado se detiene a señalizar y coopera al menos una vez con el resto de los agentes. Esto se traduce en una pérdida de tiempo en emitir señales que son aprovechadas por sus competidores. En el gráfico de la derecha se representa la magnitud de las diferencias entre ambas alternativas de comportamiento cooperativo. Se observa que, excepto en el caso particular de un sólo agente cooperativo, la ventaja del egoísmo se reduce progresivamente a medida que la

duración de los ciclos es mayor. La máxima ventaja de la reciprocidad se alcanza en el extremo de duración 10000 pasos.

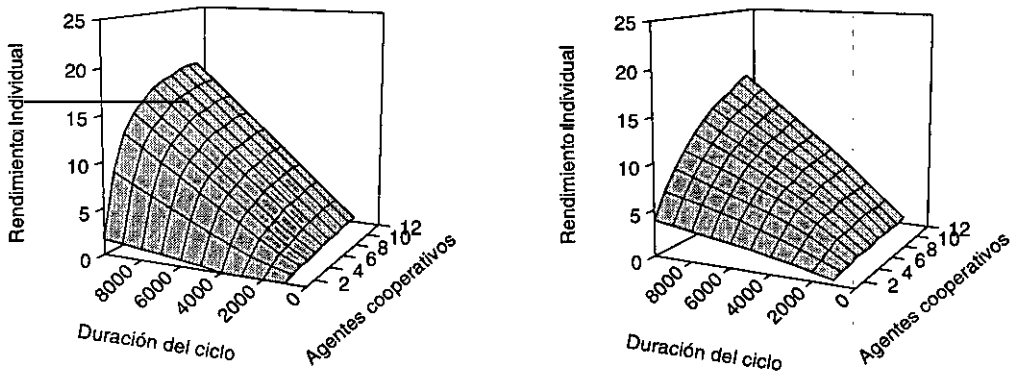


Figura 6.8. Rendimiento medio de un agente recíproco (izquierda) y un egoísta (derecha), en 100 repeticiones, en función de la duración de los ciclos de recogida y del número de agentes cooperativos en el colectivo.

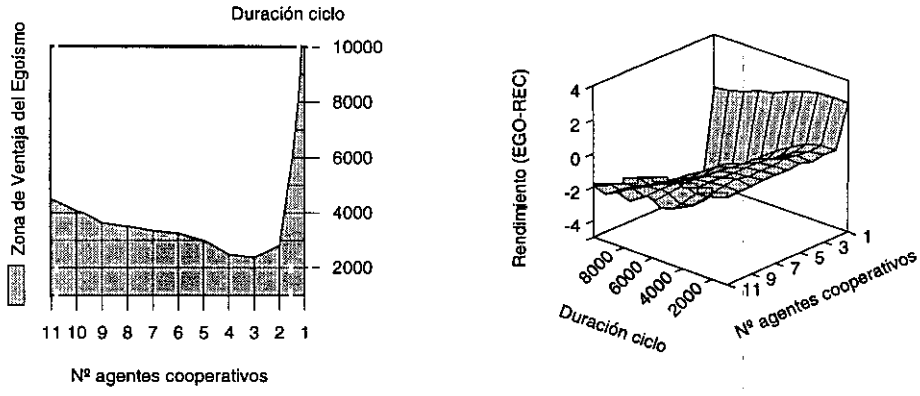


Figura 6.9. Análisis comparado del rendimiento individual de un agente recíproco y uno egoísta. En el panel de la izquierda se presenta, de forma cualitativa, el área donde el rendimiento comparado (Egoísta – Recíproco) es favorable al egoísmo. En el panel de la derecha representa cuantitativamente este rendimiento comparado.

6.2.2 Influencia de la viscosidad sobre la estabilidad de la reciprocidad: Experimentos con aprendizaje.

Los experimentos con aprendizaje se realizan con el mismo protocolo que el utilizado en los experimentos de error en el reconocimiento (sección 6.1.2) Se han probado 2 duraciones de ciclo de recogida distintos (2000 y 6000 pasos) en un sistema multiagente con 12 agentes. Estos pueden mostrar dos alternativas de comportamiento cooperativo: egoísta y recíproca. La inicialización de las tendencias a cada una de ellas se realiza de forma inespecífica y de forma adversa. En la primera, el agente decide su comportamiento inicial de forma aleatoria. En la segunda, los agentes son inicializados en el comportamiento contrario al estabilizado en la inicialización inespecífica. Las variables de tendencia inicial para $i \in \{1,2,3,\ldots,12\}$ son, $s_e(i,0)=s_r(i,0)=k_s/2$ para el caso de inicialización *inespecífica*, y $s_j(i,0)=0$, $s_k(i,0)=k_s$, $j,k \in \{$egoísta, recíproco$\}$ para la inicialización *adversa*, siendo j la alternativa estabilizada en los experimentos con inicialización inespecífica y k la contraria.

6.2.2.1 Resultados del aprendizaje con inicialización inespecífica.

En la Figura 6.10, el colectivo de agentes comienza el aprendizaje desde una situación inespecífica, resultando en una misma proporción inicial de ambas alternativas de comportamiento (egoísta y recíproco) en el colectivo de agentes. Con una duración del ciclo de 2000 pasos (panel superior), se observa que los agentes seleccionan la alternativa egoísta, desplazando al comportamiento recíproco de la población. Esto es así debido al mayor rendimiento relativo de la alternativa egoísta frente a la recíproca. Cuando un agente encuentra un objeto, si presenta el comportamiento egoísta, lo recoge y lleva al almacén partiendo de nuevo en busca de nuevos objetos. Por el contrario, si el agente presenta la alternativa recíproca, se detiene a señalizar la posición al resto de agentes. Esta señal es aprovechada tanto por los agentes cooperadores como por los egoístas que todavía no han sido catalogados como tales. Como la duración del ciclo es muy corta, no hay tiempo suficiente para identificar correctamente a los agentes egoístas ni para conseguir que los agentes recíprocos compensen, por un lado el tiempo invertido en la señalización, y por otro la ventaja adquirida por los egoístas proveniente de la primera cooperación de la estrategia TFT. Al final del aprendizaje, los agentes seleccionan la estrategia egoísta y se obtiene un rendimiento global por debajo del obtenido en la inicialización. Es importante señalar que la implantación del egoísmo en el sistema se realiza muy lentamente pues la magnitud de las diferencias entre ambas estrategias es muy pequeña con lo que determina una magnitud de los cambios en las variables de tendencia también pequeña. Los resultados de la prueba de rangos con signo de la Tabla 6.7 muestran que la pérdida en

rendimiento global entre las fases inicial y final del aprendizaje es estadísticamente significativa. Se muestra también la proporción de agentes en cada alternativa de comportamiento y un intervalo de confianza para la proporción de agentes recíprocos ilustrando la gran afinidad de los agentes hacia el egoísmo.

Tabla 6.7. Resultados obtenidos en 100 repeticiones con una duración del ciclo de 2000 pasos e inicialización inespecífica. Se muestran lo valores medios y desviaciones típicas del rendimiento global (R_g) al principio y al final de la fase de aprendizaje y el promedio de agentes en cada comportamiento al final del aprendizaje.

	Rendimiento Global		N° medio de agentes por comportamiento	
	$\bar{x}$	s	Egoísta	Recíproco
Inicial	27,29	13,28	10,24	1,76
Final	13,13	10,32		
	Z = -6,45	p<0,01	I.C(99%) $p_{recíproco}$ = 0,063±0,018	

Por el contrario, cuando la duración del ciclo de recogida es de 6000 pasos, (Figura 6.10-inferior), los agentes recíprocos tienen tiempo suficiente para reconocer a todos los egoístas y para compensar tanto el gasto de la cooperación como la ventaja inicial de los individuos egoístas. Durante el aprendizaje, los agentes seleccionan la alternativa de comportamiento recíproca sustituyendo a la egoísta produciéndose un incremento en el rendimiento global. En la Tabla 6.8 se observa el incremento estadísticamente significativo del rendimiento global entre las fases inicial y final del aprendizaje. Se muestran también la gran afinidad que presentan los agentes hacia una alternativa de comportamiento al final del aprendizaje

Tabla 6.8. Resultados obtenidos en 100 repeticiones con una duración del ciclo de 6000 pasos e inicialización inespecífica. Se muestran lo valores medios y desviaciones típicas del rendimiento global (R_g) al principio y al final de la fase de aprendizaje y el promedio de agentes en cada comportamiento al final del aprendizaje.

	Rendimiento Global		N° medio de agentes por comportamiento	
	$\bar{x}$	s	Egoísta	Recíproco
Inicial	105,40	16,97	0,54	11,46
Final	128,76	8,96		
	Z = -7,84	p<0,01	I.C(99%) $p_{recíproco}$ = 0,955±0,015	

La visión comparada de las proporciones de agentes cooperativos, y de su estimación por intervalos, en las dos situaciones de viscosidad señala la influencia de la misma en la selección de la alternativa de comportamiento.

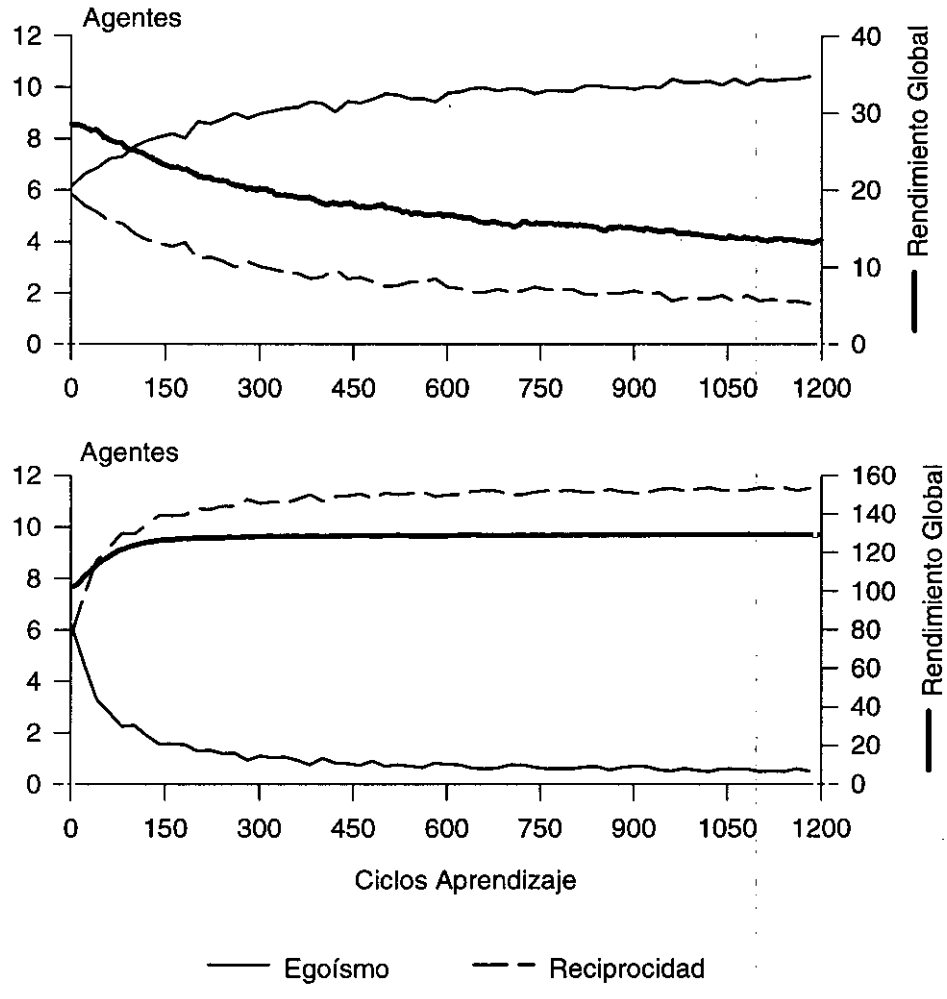


Figura 6.10. Inicialización inespecífica. Resultados de los experimentos con aprendizaje en función de la duración de los ciclos de recogida (2000 pasos en el panel superior y 6000 pasos en el panel inferior). Se muestran los valores medios sobre 100 repeticiones, del rendimiento global y del número de agentes mostrando cada una de las dos alternativas de comportamiento (egoísta y recíproca).

6.2.2.2 Resultados del aprendizaje con inicialización adversa.

En la Figura 6.11 se presentan los resultados del aprendizaje cuando el colectivo de agentes es inicializado en la situación contraria a la que se estabiliza en los experimentos de inicialización inespecífica. Cuando la duración del ciclo es de 2000 pasos (panel superior), todos los agentes comienzan mostrando el comportamiento recíproco. Esto se traduce en que el colectivo de agentes obtiene un rendimiento global cercano al óptimo calculado en las pruebas sin aprendizaje (Tabla 6.6). Sin embargo, el rendimiento individual de un agente egoísta en esa configuración de comportamientos es superior respecto al rendimiento de un agente recíproco con lo que se produce la invasión por parte de los agentes egoístas. Esta invasión trae consigo una pérdida de la eficacia colectiva. En la Tabla 6.9 se muestran los resultados medios sobre 100 repeticiones del rendimiento global antes y después del aprendizaje así como el promedio de agentes en cada comportamiento al final del mismo. La prueba de rangos con signo señala una pérdida estadísticamente significativa del rendimiento global fruto de la invasión egoísta. Esta invasión se manifiesta de también de forma notable.

Tabla 6.9. Resultados obtenidos en 100 repeticiones con una duración del ciclo de 2000 pasos e inicialización adversa en reciprocidad. Se muestran lo valores medios y desviaciones típicas del rendimiento global (R_g) al principio y al final de la fase de aprendizaje y el promedio de agentes en cada comportamiento al final del aprendizaje.

	Rendimiento Global		N° medio de agentes por comportamiento	
	$\overline{x}$	s	Egoísta	Recíproco
Inicial	36,25	8,34	10,16	1,84
Final	14,57	11,97		
	Z = -8,14	p<0,01	I.C(99%) $p_{recíproco}$ = 0,07±0,018	

Sin embargo, la situación cambia cuando el ciclo de recogida dura 6000 pasos (Figura 6.11 inferior). En este caso, los agentes permanecen un tiempo mostrando la alternativa egoísta, pues los cambios de comportamiento aislados de un agente para mostrar la alternativa recíproca se ven penalizados ya que no compensan el gasto de la cooperación al no haber más agentes recíprocos. Este problema de viabilidad inicial de la reciprocidad es similar al

que se presentaba en la sección 6.1. Sin embargo, el algoritmo de aprendizaje facilita la exploración si tras algunos ciclos exploratorios se obtienen beneficios del nuevo comportamiento explorado. Así, si se produce un cambio simultáneo de al menos 2 agentes, éstos incrementan en gran medida su rendimiento al compensarse mutuamente el gasto de la cooperación. Tras la aparición de los primeros agentes cooperativos en el equipo, la invasión por el resto es más fácil y procede de forma más rápida. Esta sustitución de la alternativa de comportamiento egoísta por la recíproca provoca un incremento en el rendimiento global. En la Tabla 6.10 se muestra los resultados medios sobre 100 repeticiones del rendimiento global antes y después de la fase de aprendizaje así como el promedio de comportamientos encontrado al final del aprendizaje. Se observan diferencias estadísticamente significativas en el cambio de rendimiento global (prueba de rangos con signo). A partir de la estimación por intervalos de la proporción de agentes recíprocos se observa que los agentes seleccionan con gran afinidad el comportamiento recíproco al final del aprendizaje.

Tabla 6.10. Resultados obtenidos en 100 repeticiones con una duración del ciclo de 6000 pasos e inicialización adversa en egoísmo. Se muestran lo valores medios y desviaciones típicas del rendimiento global (R_g) al principio y al final de la fase de aprendizaje y el promedio de agentes en cada comportamiento al final del aprendizaje.

	Rendimiento Global		Nº medio de agentes por comportamiento	
	$\bar{x}$	s	Egoísta	Recíproco
Inicial	24,40	11,64	0,71	11,29
Final	128,83	9,77		
	$Z = -8,68$	$p<0,01$	I.C(99%) $p_{recíproco} = 0,941 \pm 0,017$	

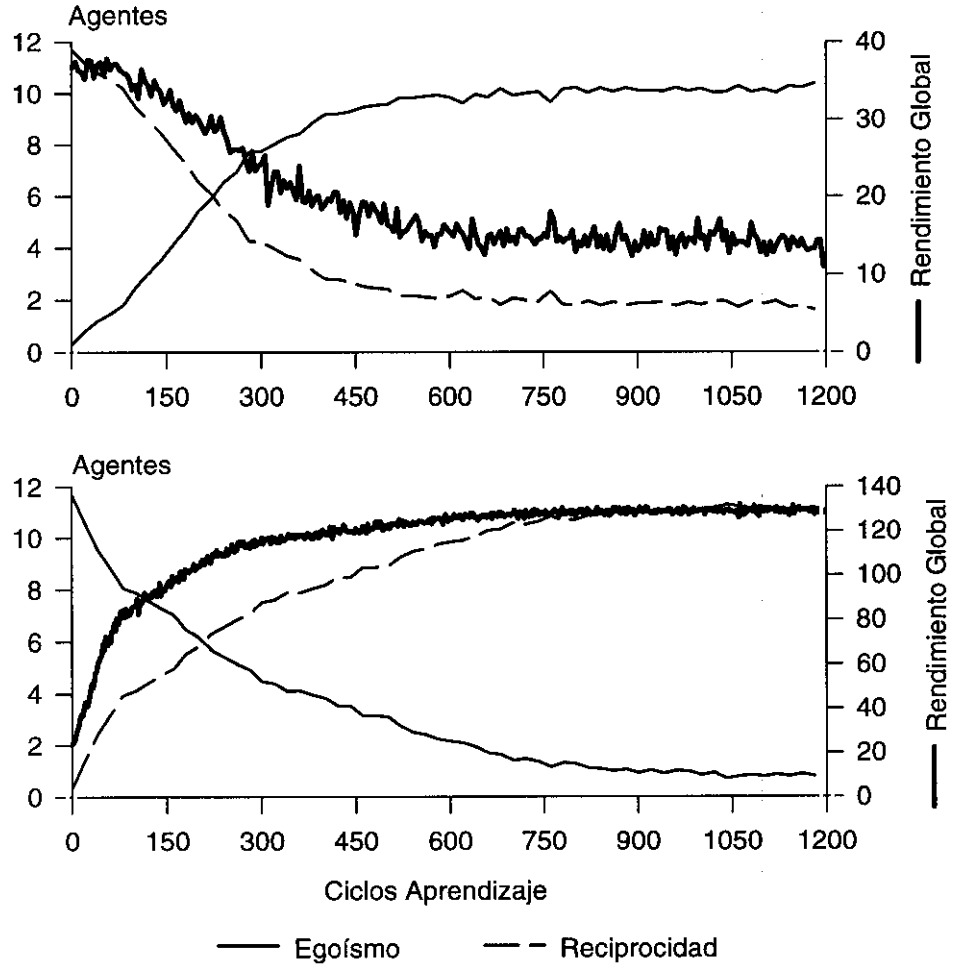


Figura 6.11. Inicialización adversa. Resultados de los experimentos con aprendizaje en función de la duración de los ciclos de recogida (2000 en el panel superior y 6000 en el panel inferior). Se muestra los valores medios sobre 100 repeticiones, del rendimiento global y del número de agentes mostrando cada una de las dos alternativas de comportamiento (egoísta y recíproca).

6.2.3 Discusión.

Los resultados de los experimentos anteriores demuestran que para la estabilización de la estrategia de reciprocidad, es necesario que los agentes dispongan de tiempo suficiente entre iteraciones del aprendizaje, es decir, suficiente duración del ciclo de recogida de objetos, que permita a los agentes recíprocos resolver tres problemas:

1. Deben catalogar correctamente al resto de los agentes según sus alternativas de comportamiento cooperativo para evitar así la explotación por parte de los agentes egoístas.
2. Deben compensar los costes de la cooperación y por lo tanto recibir la ayuda del resto de agentes cooperativos.
3. Deben recuperar la ventaja obtenida por los agentes egoístas de la primera interacción incondicional de su estrategia TFT.

Cuando la duración del ciclo de recogida es escasa (por ejemplo 2000 pasos), los agentes con comportamiento egoísta recogen objetos sin detenerse y reciben la ayuda de la primera cooperación de la estrategia TFT. En cambio, los agentes mostrando la alternativa recíproca, ante la presencia de un objeto se detienen a señalizar, no teniendo luego suficiente tiempo de recibir compensación por sus ayudas, ni tiempo para remontar la ventaja que han obtenido los agentes egoístas fruto de su primera cooperación incondicional. Esto provoca que el rendimiento individual de los agentes egoístas supere al obtenido por los recíprocos y que aquellos logren imponerse en la población.

Cuando la duración de los ciclos se incrementa (por ejemplo hasta 6000 pasos), los agentes recíprocos tienen tiempo suficiente para identificar a los egoístas como tales y evitar así cooperar con ellos, sufriendo éstos una pérdida en su tasa de recogida respecto a la que presentaban en los primeros pasos del ciclo de recogida. En cambio, los agentes recíprocos incrementan su rendimiento individual pues disponen de más tiempo para acumular los beneficios de la cooperación. Adicionalmente, cuanto mayor es la duración del ciclo de recogida, menor importancia tiene la ventaja inicial que han obtenido los agentes egoístas por la primera interacción de ayuda proporcionada por los agentes recíprocos. Como consecuencia, el rendimiento individual de los agentes recíprocos supera al rendimiento de los agentes egoístas siendo entonces posible, mediante el algoritmo de aprendizaje propuesto, establecer la cooperación recíproca en el sistema multiagente y estabilizarla frente a invasiones por parte de la estrategia no cooperativa.

6.3 Ventaja de la cooperación.

La última condición necesaria para la estabilización de la cooperación mediante reciprocidad dispone que el coste de cooperar debe ser inferior a la ventaja que supone para un individuo recibir ayuda. En la naturaleza, se encuentra que los comportamientos considerados como altruismo recíproco presentan esta relación favorable a la cooperación. Por ejemplo, en el comportamiento regurgitador de sangre en los vampiros (*Desmodus rotundus*), el incremento en el tiempo de supervivencia que obtiene un individuo que recibe ayuda es muy superior a la pérdida en que incurre el individuo que realiza la regurgitación (Wilkinson , 1984). Esta ventaja de la cooperación representa la esperanza para el individuo altruista de que sus inversiones actuales en ayudar al resto sean recompensadas en el futuro por las ayudas que recibirá.

En el sistema multiagente desarrollando la tarea de recogida de forma cooperativa, el beneficio obtenido por recibir señales de ayuda se puede cuantificar como la diferencia en el tiempo que un agente tarda en recoger un objeto y llevarlo al almacén por sí sólo y el tiempo que invertiría si lo recogiese con ayuda de las señales de ayuda. Llamemos al primer tiempo t_1 y t_2 al segundo. Por otro lado el coste que supone para un individuo la cooperación consiste en el tiempo que permanece detenido respondiendo a las peticiones de ayuda hasta que es relevado. Sea este tiempo t_3.

En el ambiente en el que se han realizado los experimentos de este capítulo, los objetos están agrupados en zonas poco accesibles (ambiente A del capítulo 5). En este ambiente, el tiempo invertido en recoger un objeto sin ayuda (t_1) es muy alto. En cambio, el tiempo invertido utilizando señales de cooperación (t_2) decrece drásticamente. Se tiene por tanto que $t_1 >> t_2$. La ventaja de la cooperación ($V_c = t_1 - t_2$), cuantificada como la diferencia entre ambos tiempos es muy alta para el ambiente descrito. Por otro lado, el coste de la cooperación (t_3) es variable en función del número de agentes cooperativos en el sistema, decreciendo a medida que aumenta este número. En general, si un colectivo tiene n agentes recíprocos, a partir del primer objeto encontrado por un agente recíproco, siempre estará uno de ellos detenido señalizando y $n-1$ agentes recogiendo objetos, por lo que en promedio, el tiempo detenido señalizando (t_3) se corresponde con $t_2/(n-1)$. Siendo esta la definición de costes y beneficios de la reciprocidad, en el ambiente A, la cooperación es ventajosa si $V_c > t_3$, es decir si $t_1 > t_2[1+1/(n-1)]$, siendo n el número de agentes cooperativos. Como los beneficios obtenidos de la cooperación son superiores a los costes se produce un incremento neto positivo en el rendimiento individual de los agentes recíprocos y por lo tanto en el rendimiento global.

En el ambiente donde los objetos están distribuidos uniformemente por el mundo de trabajo (ambiente B del capítulo 5), la relación de tiempos es distinta. La diferencia entre localizar un objeto utilizando señales de ayuda y sin su utilización disminuye, pues los

objetos son muy accesibles. Por tanto, $t_1 \cong t_2$, siendo $V_c \cong 0$. Sin embargo, el coste de la cooperación (el tiempo entre relevos) se mantiene de la misma forma que en el ambiente anterior, es decir, para un caso de dos agentes, $t_3 \cong t_2$. En esta situación desaparece la ventaja de la cooperación manteniéndose el coste de la misma por lo que ($V_c < t_3$). Por esta razón, el rendimiento individual de los agentes en un colectivo con todos ellos mostrando la alternativa de comportamiento recíproca, no aumenta al no recibir beneficios de la cooperación mientras que se paga un coste por detenerse a señalizar.

Las pruebas experimentales acerca de la ventaja de la reciprocidad se obtienen de las simulaciones con aprendizaje del capítulo 5. En el caso del ambiente A (objetos agrupados) el comportamiento recíproco se estabiliza en el colectivo de agentes, produciendo un rendimiento global cercano al óptimo. En ese ambiente, el rendimiento global es mayor cuando los agentes muestran la estrategia recíproca que cuando muestran la alternativa egoísta, por lo que el beneficio obtenido de la cooperación es superior al coste de mostrarla. En los experimentos en el ambiente B (objetos uniformemente distribuidos) durante el aprendizaje desaparecen de la población los agentes recíprocos sustituidos por la estrategia egoísta. En esa situación el rendimiento global es superior cuando todos los agentes son egoístas pues, si bien no se obtienen beneficios de la cooperación, sí desaparecen los costes de la misma.

6.4 Discusión.

En la sección anterior, se ha realizado una caracterización de los costes y beneficios del altruismo recíproco (ventaja de la cooperación) en términos de los tiempos invertidos por los agentes cooperativos en las dos tareas fundamentales que realizan, estar parados señalizando y recoger objetos.

Para el análisis de la estabilidad de la cooperación recíproca en función de la ventaja de la misma se debe obtener el rendimiento individual de las dos alternativas de comportamiento que se enfrentan. Para el cómputo de estos rendimientos, podemos utilizar las definiciones anteriores de tiempos t_1, t_2 y t_3. Para una duración del ciclo de recogida suficientemente amplia (y evitar así efectos de otros factores como la cooperación inicial de la estrategia TFT, tiempo en alcanzar el primer objeto, etc.) y ausencia de errores en el reconocimiento, el rendimiento individual de un agente es la duración total del ciclo de recogida (k_{t_limite}) dividido por el tiempo invertido por cada objeto. El tiempo por objeto utilizado por un agente recíproco (T_r) en un colectivo de m agentes donde todos los agentes son recíprocos es t_2+t_3 mientras que el tiempo por objeto de un agente egoísta (T_e) en esa configuración es t_1. De acuerdo a las condiciones de estabilidad expuestas anteriormente, la reciprocidad es

evolutivamente estable si $R_r > R_e$, o lo que es lo mismo, si $T_r < T_e$. Esta desigualdad es cierta en los ambientes donde los objetos están agrupados (ambiente A). En ellos, la reciprocidad se impondrá en la población conduciendo al sistema a presentar rendimientos globales cercanos al óptimo.

En el análisis del efecto del error en el reconocimiento (sección 6.1), hemos comprobado empíricamente que por encima de una proporción p de error, el rendimiento individual de los agentes egoístas es superior al correspondiente de los individuos recíprocos. Como resultado se tiene una invasión de la estrategia egoísta sobre la recíproca. Se produce un detrimento del rendimiento global, pues aunque en el ambiente donde se desarrollan los experimentos (ambiente A, con objetos agrupados), la ventaja de la cooperación supera a los costes de cooperar, la reciprocidad no es evolutivamente estable y el sistema termina con una configuración de comportamientos en la que lo mejor que puede hacer un agente individualmente es comportarse de forma egoísta aunque ello no se corresponda con lo mejor para el colectivo. Es importante señalar que aunque en este ambiente A, la cooperación es ventajosa, la invasión que se produce por la estrategia egoísta se traduce en que el rendimiento final de un agente egoísta es menor del que hubiera obtenido si hubiese continuado asumiendo los costes de la cooperación, que finalmente, hubiesen sido compensados por los beneficios de la misma.

El efecto que el error en el reconocimiento tiene sobre los rendimientos individuales se observa exclusivamente en el caso de los agentes egoístas. En ausencia de efectos de competencia, el rendimiento individual de los agentes recíprocos no se ve afectado por el nivel de error en el reconocimiento. Suponiendo un ambiente en el que la cooperación es ventajosa (ambiente A) y un horizonte de tiempo suficientemente grande para hacer despreciables tanto el efecto de la primera cooperación de los agentes recíprocos como el efecto del tiempo invertido en la localización del primer objeto, tenemos que el rendimiento de un agente recíproco (R_r) en ausencia de error en el reconocimiento (p=0) se corresponde con el cociente $k_{t\_limite} / (t_2+t_3)$. El rendimiento en esa misma situación para un agente egoísta es menor y se corresponde con k_{t_limite} / t_1. En la comparación de ambos rendimientos se tiene, como hemos visto, que $R_r > R_e$, dado que en el ambiente A, existe ventaja de la cooperación ya que se cumple la desigualdad $t_1 > t_2+t_3$, con lo que la reciprocidad es evolutivamente estable. El algoritmo de aprendizaje es capaz de llevar al sistema a esta situación de estabilidad obteniéndose un rendimiento global cercano al óptimo. En cambio cuando la proporción de error en el reconocimiento es extrema (p=1), el rendimiento individual de los agentes recíprocos es el mismo que en ausencia de error pero es superado por el rendimiento individual de los agentes egoístas dado que éste es ahora (k_{t_limite} / t_2) ya que recibe ayuda de todos los agentes que yerran en su reconocimiento sin pagar los costes de la cooperación. En esta nueva situación, la reciprocidad no es evolutivamente estable y es invadida por el egoísmo. Mediante el algoritmo de aprendizaje se llega a esta situación donde todos los agentes muestran el comportamiento egoísta haciendo que el rendimiento global descienda drásticamente.

En el párrafo precedente se ha asumido una duración suficiente de los ciclos de recogida para hacer despreciables tanto los efectos que sobre los rendimientos individuales de los agentes tiene el tiempo tardado hasta localizar el primer objeto por un agente recíproco, como los efectos de la cooperación obligada de la estrategia TFT con los agentes que se encuentra por primera vez. Si la duración del ciclo de recogida es menor, estos efectos dejan de ser despreciables teniendo gran influencia sobre los rendimientos individuales de ambos comportamientos. Estos efectos se engloban en lo que hemos denominado *viscosidad* de la población.

Supongamos un ambiente donde la cooperación es ventajosa (ambiente A), existe un perfecto reconocimiento entre agentes y la duración del ciclo de recogida es $k_{t\_límite}$ y sea t_i el tiempo tardado en localizar el primer objeto por un agente recíproco. La duración de un ciclo de recogida se puede descomponer en tres periodos que se corresponde con el tiempo invertido en localizar el primer objeto por un agente recíproco (t_i), el tiempo durante el que los agentes recíprocos no han identificado a los agentes egoístas y por tanto cooperan con ellos ($t_{reconocimiento}$) y, por último, el tiempo restante en el que los agentes egoístas ya han sido reconocidos (t_{resto}). La tasa de recogida de un agente recíproco durante esos tres periodos es 0 en el primer periodo (t_i), y $1/(t_2+t_3)$ en el resto ($t_{reconocimiento}+ t_{resto} = k_{t\_límite}- t_i$). En cuanto a la tasa de recogida de los agentes egoístas, ésta se corresponde en los distintos periodos con, $1/t_1$ en el primer periodo (t_i) , $1/t_2$ durante el segundo periodo ($t_{reconocimiento}$) y nuevamente $1/t_1$ en el periodo restante (t_{resto}). De la comparación de los rendimientos individuales de ambos comportamientos se tiene que los agentes egoístas obtienen ventaja sobre los recíprocos durante los dos primeros periodos de tiempo dado que sus tasas de recogida son mayores ($1/t_1>0$ para el primer periodo y $1/t_2>1/(t_2+t_3)$ en el segundo[2]). En consecuencia, ciclos de duración inferior o igual a $t_i+t_{reconocimiento}$ proporcionan ventaja al comportamiento egoísta sobre el recíproco. En esa situación, la cooperación desaparece de la población produciéndose un descenso en el rendimiento global debido a que se pierden los beneficios de la cooperación que eran superiores a los costes de la misma. Por lo tanto, paradójicamente, aunque los agentes eligen el mejor de los comportamientos individuales en cada caso, el rendimiento individual de un agente egoísta en ausencia de recíprocos es menor del que hubiera obtenido si hubiese continuado con el comportamiento recíproco. Sólo cuando t_{resto} es suficientemente grande para compensar la desventaja inicial de la cooperación, la reciprocidad puede ser estabilizada resultando en un rendimiento global óptimo.

[2] Aún siendo identificados como agentes egoístas, su rendimiento puede ser algo mayor al declarado pues pueden aprovechar ocasionalmente señales de ayuda emitidas entre agentes recíprocos, abundando entonces en la ventaja que presentan sobre los recíprocos.

6.5 Sumario del capítulo.

En este capítulo se analiza el paralelismo existente entre las condiciones de estabilidad de la reciprocidad en un sistema multiagente con aprendizaje y en la naturaleza.

Se analiza el efecto del error en el reconocimiento entre los agentes sobre el rendimiento global y los rendimientos individuales de agentes egoístas y recíprocos. Se muestran los resultados del aprendizaje en presencia de niveles de error en el reconocimiento que son tolerables y, por consiguiente, no afectan a la estabilidad de la reciprocidad, así como niveles desestabilizantes de la misma. Los resultados muestran que niveles de error en el reconocimiento elevados provocan la invasión de una población de agentes recíprocos por individuos egoístas.

Se estudia la influencia de la viscosidad de la población, determinada como la duración de los ciclos de recogida de objetos, sobre la estabilidad de la estrategia de reciprocidad. Se presentan los resultados de los experimentos con aprendizaje en situaciones de baja y alta viscosidad, determinándose la necesidad de un nivel mínimo de viscosidad para que la estrategia recíproca resista la invasión de la egoísta.

Se presenta un análisis de los ambientes probados a lo largo de la fase experimental de esta tesis en función de la relación coste/beneficio de la cooperación en los mismos. Se discute esta relación y su influencia sobre la estabilidad de la reciprocidad observándose que para la misma es necesario que la cooperación sea ventajosa.

Por último se discute el efecto de los factores anteriores sobre la estabilidad de la reciprocidad, justificándose la necesidad del cumplimiento de todas las condiciones anteriores para que la estrategia cooperativa recíproca sea evolutivamente estable.

Capítulo 7

Implementación física del robot COOBOT

En este capítulo se presenta la implementación realizada de la arquitectura del agente recíproco autónomo (AREA) en el robot COOBOT[1] (Murciano, A. *et al.*, 1997). Se describen las soluciones adoptadas para los dispositivos sensoriales, efectores y de control, así como una evaluación del funcionamiento de los comportamientos incluidos en el agente AREA. Finalmente se presentan los resultados de los experimentos de 2 robots COOBOT realizando la tarea de recogida de objetos planteada en la tesis.

7.1 Necesidad de la implementación hardware.

Las simulaciones realizadas en la fase experimental de esta tesis, se desarrollan sobre una representación más o menos reduccionista de las características del ambiente real. Como en toda modelización, la validez de los resultados depende en gran medida de la fiabilidad del modelo realizado. Las interacciones agentes-entorno "virtuales" pueden carecer de los

[1] El desarrollo del robot COOBOT ha sido financiado por medio del proyecto de investigación 95/5548 de la Universidad Complutense de Madrid y con fondos del Dpto. de Matemática Aplicada (Biomatemática). Esta plataforma está pendiente de patente (exp. N° 9502259)

efectos propios del trabajo en el mundo real, pues éstos son frecuentemente desconocidos o de difícil modelización. A modo de ejemplo, la simulación de las lecturas sensoriales de un agente puede no tener en cuenta los posibles errores sistemáticos debidos a una mala calibración del sensor que se simula. De la misma forma, la ejecución de una acción motora en el mundo físico frecuentemente va acompañada de procesos de difícil modelización como son la inercia, el rozamiento, e incluso los efectos que las imperfecciones del terreno tienen sobre los efectores.

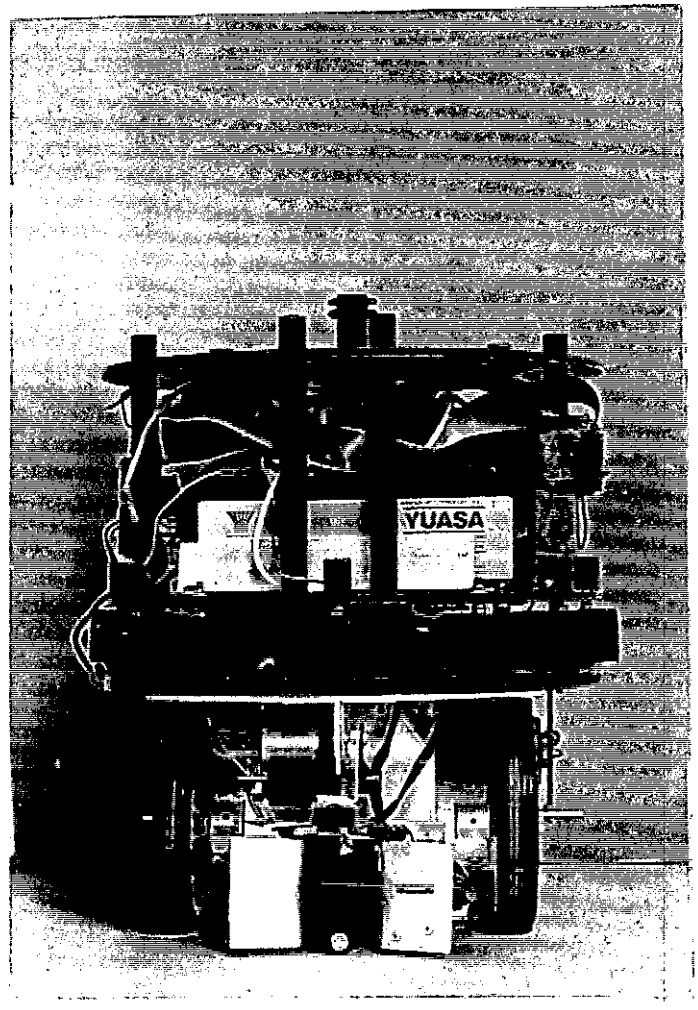
El simulador utilizado y la arquitectura de los agentes autónomos (descrita en el capítulo 4), han sido diseñados teniendo en cuenta las características y restricciones del robot físico COOBOT, desarrollado en nuestro departamento. Para incrementar la validez externa de los resultados obtenidos en simulación se ha implementado la arquitectura AREA en dicho robot móvil. Se realizan pruebas del funcionamiento de los distintos comportamientos de la arquitectura de los agentes para evaluar la adecuación de los mismos para plataformas robóticas y mostrar la similitud entre los resultados obtenidos en el robot real y los obtenidos en las simulaciones.

A continuación se describe el robot utilizado y los resultados obtenidos en las pruebas de evaluación de los distintos comportamientos.

7.2 Descripción del robot COOBOT.

El robot es una plataforma móvil de planta circular de 30 cm de diámetro y una altura de 50 cm. (Fotografía 1). Cuenta con 2 baterías conectadas en paralelo de 6 V (4 Ah) dedicadas a la alimentación de la unidad de proceso, dispositivos sensoriales y motores, y con 1 batería de 12 V (1.4 Ah) para la alimentación independiente del sistema de comunicación. Su autonomía de funcionamiento es aproximadamente de 1 hora con un tiempo de recarga de 4 h. El peso total aproximado en orden de marcha es de 10 Kg.

El ambiente donde trabaja el robot (Fotografía 2) es un espacio de unos 20 m^2 rodeado de paneles obscuros e iluminado con luz blanca artificial. Los paneles pueden disponerse en distintas configuraciones creando pasillos, obstáculos, etc. cuya disposición final determinará la arena de trabajo. Los robots deben recoger una serie de objetos presentes en el ambiente. Se dispone una zona de almacén donde deben depositarse los objetos capturados. Este almacén emite una señal característica que los robots utilizan para su localización.



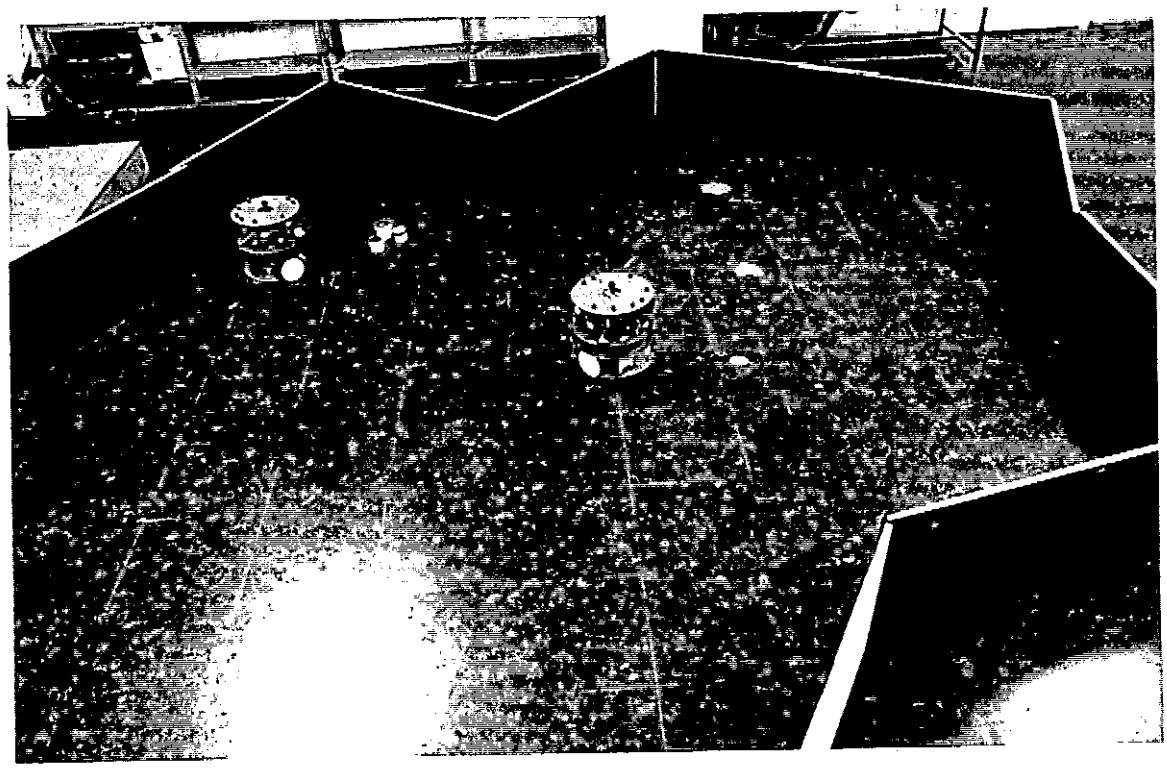
Fotografía 1. El robot COOBOT.

Los distintos aspectos funcionales del robot se agrupan en 4 bloques fundamentales dedicados a la percepción del mundo, la comunicación, la interacción motora con el ambiente y la unidad de control.

7.2.1 Dispositivos sensoriales.

El robot cuenta con dos tipos de sensores distintos que se combinan en la adquisición de la información necesaria para la ejecución de los comportamientos de la arquitectura AREA. Estos son sensores de luz (que cuantifican cantidad de luz presente en su entorno local) y sensores de contacto. COOBOT tiene una estructura física tal que permite fácilmente los

cambios de orientación y posición de dichos sensores. A continuación se enumeran estos dispositivos sensoriales según su relación con distintos aspectos funcionales del robot.



Fotografía 2. El ambiente de trabajo.

Navegación.

La información que el robot utiliza para la navegación es obtenida de forma pasiva mediante los dos tipos de sensores.

- Fotorresistencias (LDR).

  Estos sensores responden a la cantidad de luz que captan en la zona hacia donde se orienta el sensor, de forma aproximadamente lineal en el rango de 0.2 a 1 metro de los obstáculos. Están situados en la parte frontal superior de la plataforma móvil y espaciados 30° entre ellos. Dada su sensibilidad a la luz ambiente, están protegidos en el interior de cilindros de color negro que evitan la incidencia directa de los focos de iluminación. Las lecturas de estos sensores proporcionan una estimación de la distancia

hacia los obstáculos, de forma que a medida que el robot se aproxima a ellos, los sensores más próximos al obstáculo disminuyen su nivel de activación.

- Microinterruptores (Bumpers).

  Estos sensores están localizados en las mismas posiciones que las fotorresistencias. Responden al contacto con los obstáculos de forma que un choque del robot activa el bumber (o los bumbers) de la zona de choque. Para la seguridad del robot, la activación de estos sensores es integrada en el comportamiento de navegación anulando el movimiento que provocó el contacto y desplazando al robot en la dirección opuesta a la zona de contacto.

Percepción de objetos.

Los objetos que deben ser recogidos tienen la superficie superior metálica, son de forma cilíndrica y recubiertos de un papel reflectante para facilitar su localización. La detección de los objetos se realiza activamente mediante los siguientes dispositivos.

- Sensor de objetos

  Este sensor es una fotorresistencia (LDR) situada en la base del robot, cercana al suelo y orientada hacia adelante. El sensor está combinado con un emisor de luz focalizado. La reflexión del haz de luz sobre la superficie reflectante del objeto activa este sensor. El rango de percepción de los objetos es aproximadamente de unos 50 cm.

- Sensor de contacto de objetos.

  Este sensor está situado en la misma zona que el anterior y se activa mediante el contacto con un objeto. La activación del microinterruptor determina que el objeto está dentro de la zona de captura. Existe otro bumber a distinta altura que el anterior para la discriminación entre objetos de dos alturas diferentes (en los experimentos de esta tesis no se hace uso de él).

Comunicación.

La comunicación entre robots se realiza mediante una corona de 8 emisores y 8 receptores de señales de infrarrojos espaciados 45° entre sí. Esta comunicación es omnidireccional en la emisión, pero mediante una cuantificación de la intensidad de la señal en los distintos receptores de la corona, el robot receptor puede determinar la dirección relativa del emisor. Una serie de relés se encarga del encendido y apagado de este sistema así como la conmutación entre los modos de emisión/recepción. La comunicación es seriada con una velocidad máxima de transmisión de 1200 bps. El alcance de las señales es aproximadamente de unos 3 metros, atenuándose esta señal con la distancia. Los

obstáculos impiden la propagación de las señales. El sistema de comunicación utiliza un lenguaje compacto de mensajes de un byte de tamaño. En un mismo mensaje pueden incluirse los tipos de mensaje enumerados en la sección 4.1.3. y la identificación del robot emisor. El almacén emite una señal característica que es utilizada por el robot para su localización.

7.2.2 Dispositivos efectores.

Existen dos dispositivos efectores distintos dedicados a la navegación y a la captura de objetos respectivamente.

El dispositivo motor para la navegación, se ha implementado mediante dos ruedas motrices conectadas, mediante un reductor mecánico, a sendos motores DC de alto par y una velocidad en carga aproximada de 35 cm/s. La activación de ambos motores se decide mediante un sistema electrónico que permite su control bidireccional. Este sistema proporciona los comandos motores descritos en la Tabla 4.1.

El dispositivo efector de recogida de objetos se resuelve mediante un sistema magnético acoplado a una grúa con un motor paso-a-paso que efectúa movimientos de subida y bajada. Existen tres posiciones de este dispositivo que determinan, por un lado la altura de captura del objeto, la elevación del mismo a la posición de transporte y finalmente la elevación de la grúa a la posición de reposo. En el trayecto de la grúa hacia la posición de reposo provoca que el objeto transportado choque con un tope y se desprenda del soporte magnético.

7.2.3 Dispositivo de control.

El control del robot descansa sobre una tarjeta de PC con un procesador NEC V25, compatible 100% con la familia de procesadores 80x86, lo que facilita su programación mediante herramientas estándar. Su frecuencia de procesamiento es de 16 MHz. Es de reducido tamaño (10.7 x 9.2 cm.) y dispone de 512 Kb de memoria RAM y 256 Kb de memoria flash permanente. Dispone de 2 puertos de comunicación serie, utilizados como consola y como conexión con el sistema de comunicación, y 2 puertos paralelos utilizados como dispositivo de entrada/salida para el control del resto de dispositivos del robot. La información recogida por los distintos dispositivos sensoriales es digitalizada mediante un conversor analógico/digital de 19 canales y de gran velocidad de conversión. El resultado de las conversiones es utilizado por el algoritmo de control para la ejecución de los distintos comportamientos.

El algoritmo de control del robot es el mismo que el utilizado por los agentes en el simulador. Una vez compilado se trasvasa a la tarjeta controladora mediante el puerto serie de la consola, grabándose en la memoria flash. El encendido del robot supone la ejecución del programa de control de forma autónoma, aunque el robot puede ser monitorizado mediante un ordenador externo conectado a la consola.

Es importante señalar que la elección de una arquitectura reactiva basada en comportamientos, como la arquitectura AREA, conduce a que el funcionamiento de los agentes, físicos y virtuales, no dependa críticamente de la exactitud tanto de la percepciones sensoriales como de los resultados de las acciones motoras. La ausencia de procesos de planificación, utilización de mapas globales o procesos deliberativos y negociadores, reduce drásticamente la importancia futura de errores de percepción o acción del robot. Estas características facilita la portabilidad de la arquitectura entre las dos plataformas probadas.

7.3 Evaluación del funcionamiento.

Para la evaluación de los distintos comportamientos en el robot utilizaremos criterios cualitativos como los sugeridos en (Colombetti, Dorigo y Borghi, 1996). Estos se refieren a la *adecuación* y la *robustez* del comportamiento. Un comportamiento es adecuado si cumple el objetivo para el que está diseñado en las condiciones ambientales especificadas en su diseño. El comportamiento es además robusto, si cumple con su objetivo aún en condiciones distintas a las planteadas en su diseño. Por ejemplo, un comportamiento de evitación de obstáculos es adecuado si evita chocar con los obstáculos para unas condiciones de iluminación dadas. Es además robusto si continúa ejecutándose correctamente para distintas condiciones de luz. Adicionalmente, consideraremos robusto un comportamiento que siga cumpliendo su objetivo aún cuando existan errores o imperfecciones en el resultado de su ejecución. En el ejemplo del comportamiento de evitación de obstáculos, la magnitud del desplazamiento del robot mediante la ejecución de un giro depende de la carga de las baterías, de la presencia de efectos de inercia y de deslizamiento de las ruedas, etc. El comportamiento es robusto si continúa evitando colisiones con los obstáculos a pesar de estos efectos.

La evaluación de los comportamientos de la arquitectura AREA en el robot COOBOT se realiza en dos vertientes distintas. Por un lado se evalúa cada comportamiento aislado en un único robot según los criterios de adecuación y robustez planteados anteriormente. Una siguiente evaluación consiste en la realización de los experimentos de recogida de objetos de forma cooperativa por 2 robots COOBOT.

7.3.1 Evaluación de los comportamientos aislados.

Para la evaluación de los comportamientos individuales de la arquitectura AREA se realizan experimentos en los que se sitúa el robot en el mundo de trabajo frente a los estímulos desencadenantes de cada comportamiento reactivo. A continuación se desglosan los resultados de los experimentos de navegación y captura de objetos y se señala las diferencias observadas con los resultados obtenidos en los agentes virtuales.

7.3.1.1 Navegación.

Los comportamientos relacionados con la navegación son los de evitación de obstáculos y exploración. El robot parte de una posición fija del ambiente de trabajo, el cual contiene varios obstáculos (entre ellos otro robot). Comienza a navegar por el ambiente y utilizando el comportamiento de **`evitación-de-obstáculos`**, evita que se produzcan colisiones con los obstáculos. El disparo ocasional del comportamiento de **`exploración`** complementa la navegación evitando que el robot se mueva indefinidamente por una misma ruta.

El funcionamiento de la navegación del robot COOBOT es correcto. Es necesario mencionar, sin embargo, una serie de particularidades.

- El sistema sensorial para la navegación presenta dependencia del nivel de luz ambiental en el mundo de trabajo. Diferentes intensidades de iluminación provocan en el agente un comportamiento variable de aproximación a los obstáculos. Para incrementar la robustez del comportamiento de navegación y evitar esta dependencia del ambiente, se realiza un proceso de autocalibrado, en el que los sensores adaptan su lectura en función del nivel de luz ambiental.
- La detección de otro robot como un obstáculo se ve dificultada dado que el gradiente de luz alrededor de un robot es menor al que presentan el resto de los obstáculos del ambiente. Esto provoca que un robot se aproxime a menor distancia a otro robot que al resto de los obstáculos.
- La situación de los sensores de navegación en la estructura del robot provoca la existencia de zonas ciegas entre sensores a corta distancia. Por ello, un agente presenta algunas dificultades para percibir un borde de un obstáculo estrecho si éste se sitúa entre los ámbitos de percepción de dos sensores.
- La velocidad de giro de los motores del robot decrece con la carga de las baterías. Esto provoca que los movimientos de giro de COOBOT decrezcan en amplitud a lo largo del tiempo, por lo que se observan con mayor frecuencia y mayor intensidad (se incrementa la frecuencia de los giros totales -GIT y GDT- en relación con los giros parciales -GIP y GDP).

- La ejecución de una acción motora del robot está acompañada de fenómenos de inercia e imprecisión. Por ejemplo, se comprueba que la posición final que tiene un robot tras una secuencia de movimientos no es repetible. Estas deficiencias, sin embargo, no impiden el funcionamiento correcto de la evitación de los obstáculos. Por el contrario, complementan al comportamiento exploratorio de los agentes impidiendo que el agente se quede bloqueado en secuencias de comportamiento simétricas.

A pesar de las particularidades descritas, la navegación y exploración observada en el robot COOBOT demuestran ser robustas y adecuadas, siendo equivalentes a las que se observan en el modelo realizado en el simulador.

7.3.1.2 Recogida no guiada de objetos

Se deposita en el ambiente una serie de objetos en posiciones alejadas de los obstáculos y el robot parte de una posición fija en busca de los mismos. El robot detecta un objeto a una distancia aproximada entre 50 y 75 cm. y se aproxima hacia él por medio de la activación del comportamiento de **`seguimiento-de-objetos`**. Debido a que el desplazamiento del robot puede no seguir una trayectoria perfecta, el robot en ocasiones pierde la percepción del objeto activándose el comportamiento de **`pérdida-de-objetos`**. Este último desencadena los desplazamientos laterales necesarios para volver a percibir el objeto. Cuando se activa el microinterruptor de contacto de objetos se desencadena el comportamiento de **`captura-de-objetos`** bajando la grúa hasta la posición de captura. Una vez alcanzada esta posición, el objeto se adhiere al soporte magnético de la grúa que se eleva hasta la posición de transporte. A partir de este momento, trata de localizar la señal de almacén. Una vez percibida, el robot es capaz de interpretarla correctamente y dirigir la orientación de su desplazamiento hacia el lugar de donde procede.

No se observan diferencias substanciales entre los comportamientos relacionados con la recogida de objetos en el robot físico y los agentes en el simulador.

7.3.1.3 Comunicación entre agentes.

Los agentes demuestran ser capaces de realizar emisiones de mensajes interpretables por otros agentes. Son igualmente capaces de determinar la dirección del emisor del mensaje en función de la activación diferencial de los distintos sensores de su corona de receptores de infrarrojos.

Los errores de comunicación entre los robots son frecuentes (aproximadamente 50%). Entre ellos, se debe distinguir entre los errores que suponen la pérdida completa del mensaje y los que conllevan una mala interpretación del mensaje recibido. Los primeros errores son los más frecuentes, sin embargo, no afectan al reconocimiento del engaño

(discutido en la sección 6.1) pues basta con la recepción de un único mensaje de donación de ayuda o de relevo de señalizador para que los agentes se identifiquen correctamente. Este error en el reconocimiento se ha cuantificado experimentalmente situándose muy por debajo del 20%. La pérdida de señales tampoco impide la realización de la tarea de búsqueda guiada de objetos ni de acercamiento al almacén ya que, para la orientación correcta del robot hacia el emisor de la señal, basta un número reducido de señales percibidas correctamente. El otro tipo de errores provoca la recepción de mensajes ruidosos que carecen de significado pues, en su mayoría, se corresponden con tipos de mensaje no incluidos en el vocabulario de la arquitectura AREA. La consistencia de los mensajes recibidos es analizada de acuerdo a criterios de filtrado que, por un lado, exigen que el mensaje tenga significado, es decir, sea uno de los posibles mensajes utilizados por el sistema de comunicación y, por otro lado, los mensajes con significado correcto deben ser reiterados para ser tomados en cuenta. Este sistema de filtrado incrementa la robustez del sistema de comunicación.

La inherente secuencialidad de los procesos de simulación hacen que la comunicación real sea difícil de implementar de forma realista en simulación. Sin embargo, la inclusión de niveles de ruido variable en la simulación de estos procesos aproxima, en términos del resultado obtenido, la eficacia del comportamiento que los robots físicos presentan en el mundo real.

7.3.2 Experimentos de cooperación con COOBOT.

Una vez probada la adecuación y la robustez de los comportamientos de la arquitectura AREA en el robot físico, se plantea la realización de los experimentos de recogida real de objetos de forma cooperativa. El objetivo de estos experimentos es evaluar la eficacia de las tres alternativas del comportamiento cooperativo (altruista, recíproca y egoísta) en la recogida real de objetos en el mundo físico para analizar su similitud con los resultados obtenidos en simulación. Para ello, se sitúan dos robots en el ambiente de trabajo y los objetos se disponen agrupados en una zona de paso poco frecuente. En este sentido el ambiente es similar al ambiente A probado en los experimentos de simulación (Tabla 5.1). Los robots parten de la zona de almacén navegando por el ambiente en busca de los objetos. El experimento se detiene transcurridos 20 minutos y se cuantifica el número de objetos recogidos y depositados en el almacén por cada agente en las 6 combinaciones posibles de un equipo de 2 agentes con las tres alternativas de comportamiento cooperativo. Los resultados de 12 repeticiones se muestran en la Tabla 7.1, donde se representa la media y desviación típica del número de objetos recogidos por cada robot en las distintas combinaciones del equipo de trabajo.

Tabla 7.1. Resultados de los experimentos con robots reales (12 repeticiones). Se muestran los valores medios y las desviaciones típicas del número de objetos recogidos por el colectivo (R_g) y recogidos por un robot (R_a, R_e, R_r,) según la combinación de comportamientos cooperativos en un equipo de 2 agentes. Con líneas de puntos se representa la ausencia de un comportamiento en la combinación.

	R_g		R_a		R_e		R_r	
	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s
AA	10,7	2,49	5,35	1,25	----	----	----	----
AR	11,6	2,68	5,5	1,50	5,7	1,28	----	----
AE	9,35	3,37	0,0	0,0	----	----	9,35	3,37
RR	9,95	2,85	----	----	4,97	1,43	----	----
RE	6,55	2,11	----	----	2,75	1,21	3,8	1,89
EE	7,05	2,78	----	----	----	----	3,52	1,39

Al igual que los resultados sin aprendizaje de la sección 5.1, en la Tabla 7.1 se observan diferencias significativas en el rendimiento global según las distintas configuraciones de comportamientos cooperativos en el equipo de trabajo (prueba de Kruskal-Wallis $\chi^2 = 40{,}398 \quad p<0{,}01$). En el ambiente utilizado, las combinaciones de comportamientos que obtienen un rendimiento global menor son la formada por 2 robots egoístas y la formada por un recíproco y un egoísta[2]. El rendimiento global del resto de configuraciones es similar pues en todas ellas se tiene que, en promedio, un robot está recogiendo objetos utilizando señales de ayuda mientras que el otro robot está detenido emitiéndolas. La combinación AE, en la que están presentes un agente altruista y uno egoísta, obtiene un rendimiento ligeramente inferior al obtenido por dos agentes cooperativos, pues se incrementa el tiempo hasta que el individuo altruista localiza el primer objeto.

El rendimiento individual de un robot con un determinado comportamiento depende del comportamiento que presente su compañero. En este sentido, los robots cooperativos (altruistas y recíprocos) presentan su mejor rendimiento individual cuando el otro robot

[2] La configuración formada por un robot recíproco y otro egoísta (RE) es un caso particular, dado que el robot recíproco se comporta de forma egoísta cuando agota su tiempo de señalización sin haber recibido ayuda de su compañero egoísta, con lo que su rendimiento global es similar al obtenido por la combinación EE aunque algo menor pues el agente recíproco ha estado detendido un cierto tiempo hasta agotar su tiempo límite de señalización.

también es cooperativo. Por otro lado, un robot egoísta presenta su mayor rendimiento cuando el otro robot es altruista pues recibe su ayuda sin pagar ningún coste de la cooperación.

El planteamiento experimental diseñado cumple con las condiciones de estabilidad de la reciprocidad discutidas en el capítulo 6. Por un lado, el sistema de comunicación entre robots, con el análisis de consistencia y repetibilidad de los mensajes, resulta en un porcentaje de error en el reconocimiento prácticamente nulo, que como se ha demostrado es tolerable en las simulaciones de la sección 6.1. La duración de los ciclos de recogida es suficientemente alta para dotar al sistema de la viscosidad necesaria para la estabilización de la reciprocidad (sección 6.2). Esto es, durante los 20 minutos de duración del experimento se producen un número de encuentros entre robots suficiente para compensar la ventaja inicial de un robot egoísta fruto de la cooperación inicial de los estrategas TFT. Por último, el análisis de los tiempos invertidos en la localización de los objetos con y sin señales de cooperación supone que, en el ambiente dispuesto, los costes de la cooperación sean inferiores a la ventaja de ser ayudado (sección 6.3). Concretamente la estimación del tiempo medio de localización y depósito de un objeto sin mensajes de señalización es superior al tiempo utilizado en recogerlos cuando se dispone de señales cooperativas más el tiempo de espera hasta la señal de relevo.

Los resultados obtenidos en los experimentos, junto con el cumplimiento de las condiciones de estabilización de la reciprocidad por parte del planteamiento experimental utilizado hacen que se pueda afirmar que la cooperación recíproca es una estrategia estable cuando se adopta por ambos robots y que conduce al sistema a la obtención de rendimientos óptimos.

7.4 Discusión.

A pesar de las dificultades y diferencias advertidas del trabajo en entornos físicos, todos los comportamientos probados de la arquitectura AREA muestran su adecuación para el trabajo con el robot COOBOT y presentan buenos grados de robustez. En consecuencia, la implementación del sistema de control de la arquitectura AREA a la plataforma física, permite realizar de forma eficiente la tarea de recogida de objetos en todos sus aspectos, como son, la navegación y exploración del ambiente, la búsqueda independiente o guiada de objetos, la localización y captura de los mismos y el posterior transporte al almacén. Tras el análisis de los resultados de la arquitectura AREA en el robot COOBOT se puede concluir que el funcionamiento de los comportamientos de dicha arquitectura es

equiparable en ambas plataformas, contrastándose la verosimilitud de los resultados obtenidos en simulación.

Los resultados obtenidos de los experimentos de recogida cooperativa de objetos demuestran que la adopción de una estrategia cooperativa por ambos robots en el colectivo (ya sea la alternativa altruista o la recíproca) incrementa la eficacia del colectivo. La adopción por parte de los robots de la estrategia de comportamiento recíproca conduce a la estabilización de la cooperación en el colectivo y, por lo tanto, a la obtención de rendimientos óptimos.

7.5 Sumario del capítulo.

En este capítulo se describe la plataforma hardware COOBOT desarrollada en nuestro departamento y la implementación física que se ha realizado de la arquitectura AREA. Se muestran los resultados del funcionamiento de los distintos comportamientos, tanto individuales de navegación y captura de objetos, como colectivos de comunicación y seguimiento de señales, apuntando en cada caso las particularidades que se derivan de la interacción con el mundo físico. En todos los casos se muestra la adecuación de la arquitectura AREA para el trabajo con robots físicos.

Se presentan los resultados de los experimentos sin aprendizaje de captura de objetos por un equipo de dos robots COOBOT. Los resultados obtenidos permiten afirmar el paralelismo existente en el funcionamiento del sistema robótico en las plataformas física y de simulación.

Capítulo 8

Conclusiones

En esta tesis se ha presentado un modelo de cooperación en robótica colectiva inspirado en la biología, que reúne en una misma arquitectura las características de adaptabilidad mediante evaluaciones de eficacia individual y estabilidad de la cooperación altruista. La presencia simultánea de ambas características en la arquitectura de los agentes le confieren gran atractivo para el trabajo colectivo en entornos desconocidos y cambiantes. Así, se propone una nueva arquitectura de agente autónomo (AREA) que presenta los comportamientos necesarios para mostrar la estrategia de reciprocidad. Esta arquitectura incluye la capacidad de adaptación al ambiente mediante aprendizaje con señales de refuerzo locales a los agentes. El modelo presentado permite la estabilización de estrategias de cooperación altruista, que aunque son costosas para el individuo que las muestra, procuran beneficios al colectivo llevándole a alcanzar rendimientos cercanos al óptimo en los ambientes donde se desarrolle la tarea elegida.

La arquitectura AREA ha sido implementada en agentes autónomos simulados probándose en diferentes ambientes de experimentación. En todos ellos, el algoritmo de aprendizaje ha permitido la estabilización de la estrategia cooperativa óptima. Esta arquitectura ha sido también implementada en robots reales donde se ha probado la adecuación de los distintos comportamientos para el trabajo en entorno real.

En la arquitectura de los agentes, se incluye un repertorio de comportamientos reactivos que posibilitan el trabajo desde el primer momento con niveles de eficacia aceptables. Se

incluye además un repertorio de comportamientos sociales cuya eficacia depende del ambiente siendo estos comportamiento sujeto de aprendizaje. El control del sistema multiagente es descentralizado y emerge de las interacciones entre los agentes del colectivo, presentando robustez en su funcionamiento, pues resiste perturbaciones producidas por errores en el funcionamiento de algún agente o errores en la comunicación, demuestra adaptabilidad en su comportamiento, modificando la estrategia de cooperación según los requerimientos del ambiente y finalmente, alcanza optimalidad en su rendimiento pues se obtienen rendimientos cercanos al óptimo para todos los ambientes probados.

El análisis de los resultados obtenidos de los distintos experimentos, arroja las siguientes conclusiones:

- *Optimalidad dependiente del ambiente.* La eficacia de los comportamientos cooperativos altruistas depende del ambiente donde se muestran y de la frecuencia en que este presentes en el colectivo.
- *Adaptabilidad.* El algoritmo de aprendizaje mediante evaluaciones de la eficacia individual, posibilita la adaptación del comportamiento cooperativo de los agentes al ambiente donde trabaja, aun siendo desconocido a priori, permitiendo alcanzar rendimientos colectivos óptimos o cercanos al óptimo.
- *Comportamientos reactivo:* La inclusión en la arquitectura de los agentes de comportamientos reactivos posibilita su funcionamiento de forma satisfactoria desde el primer momento, haciendo posible concentrar los esfuerzos del aprendizaje sobre el comportamiento cooperativo.
- *Inestabilidad del altruismo.* El comportamiento altruista no es evolutivamente estable bajo las condiciones de aprendizaje utilizadas en el sistema multiagente pudiendo ser desplazado por alternativas de comportamiento egoístas.
- *Estabilidad de la reciprocidad.* La inclusión de la estrategia de reciprocidad en la arquitectura de los agentes, posibilita que la cooperación altruista se convierta en un comportamiento evolutivamente estable.
- *Flexibilidad.* Si las condiciones del ambiente cambian, el algoritmo de aprendizaje permite salir de la configuración de comportamientos estabilizada y seleccionar la alternativa de comportamiento que proporcione mejores rendimientos.
- *Exploración/Explotación.* El factor de exploración variable del algoritmo de aprendizaje permite explorar nuevas soluciones en busca de mejoras en el rendimiento, sin perjuicio del rendimiento global.

- *Escalabilidad.* El cambio en el número de agentes en el equipo de trabajo no afecta a las propiedades de la arquitectura de reciprocidad implementada en el sistema multiagente.
- *Competición.* La existencia de competición por los recursos incrementa los costes de la cooperación altruista. Pese a esta dificultad adicional, el sistema es capaz de aprender la estrategia óptima y estable.
- *Error de reconocimiento.* La falta de reconocimiento entre los agentes es un aspecto crítico para la estabilidad del altruismo recíproco. El sistema de reconocimiento implementado en la arquitectura de los agentes permite la estabilización del altruismos en presencia de niveles de error moderados.
- *Viscosidad de la población.* La estabilización de la reciprocidad necesita necesariamente de la existencia de suficiente viscosidad en la población. En ausencia de suficiente número de interacciones entre agentes, la reciprocidad no es estable, dando paso a comportamientos sociales egoístas. El uso de señales de refuerzo acumulado a lo largo del tiempo permite compensar los costes del altruismo facilitando el aprendizaje.
- *Ventaja de la cooperación.* La cooperación altruista sólo es estable en los ambientes donde los beneficios de la cooperación superen a los costes de la misma.
- *Adecuación a plataformas Hardware.* Los resultados de la arquitectura de reciprocidad implementada en el robot COOBOT demuestran la adecuación de la misma a trabajos en entornos reales.

La Inteligencia Artificial frecuentemente busca inspiración en la biología con la intención de extraer soluciones concretas o incluso generar nuevas metodologías de trabajo que sean de utilidad para resolver los problemas a los que se enfrenta. La idea subyacente es que la naturaleza ha solucionado de forma eficiente problemas similares a los que aborda la Inteligencia Artificial. El éxito con el que cuentan los seres vivos viene de su gran capacidad de adaptación a las distintas presiones ambientales. Mediante largos procesos de evolución, los animales han seleccionado, por un lado los comportamientos más adecuados para distintas presiones selectivas, y por otro han desarrollado las estructuras y los mecanismos necesarios para permitir la modificación del comportamiento durante la vida del individuo y así adaptarse a nuevas presiones o mejorar las soluciones heredadas.

Esta idea ha sido también inspiradora de la presente tesis. Hemos extraído de la naturaleza estrategias de cooperación que permiten unificar en una misma arquitectura dos de los aspectos fundamentales perseguidos en el trabajo con sistemas multiagente. El primero de ellos es el objetivo de conseguir rendimientos óptimos en la ejecución del trabajo. El segundo aspecto, relacionado con el anterior, es la pretensión de que los sistemas

artificiales sean capaces de modificar su comportamiento para poder alcanzar rendimientos óptimos en cualquier ambiente. La arquitectura propuesta en esta tesis incluye comportamientos cooperativos que son elegibles e incrementan el rendimiento del colectivo, y un algoritmo de aprendizaje por refuerzo que evalúa la eficacia individual del agente. La acción conjunta de ambos mecanismos soluciona el grave problema de estabilidad que surge del conflicto entre los intereses propios de un individuo que puede modificar su comportamiento, y los intereses del colectivo, permitiendo estabilizar, la cooperación altruista en los ambientes donde su presencia sea beneficiosa.

Trabajos futuros.

Las conclusiones obtenidas en esta tesis han dado lugar a una serie de interrogantes y de nuevos caminos a investigar que inspirarán los próximos trabajos que emprenderemos. En todos ellos, la biología será una fuente de inspiración, por un lado para mejorar las propuestas realizadas en la tesis, y por otro procurar soluciones a distintos problemas que existen en la robótica colectiva.

Mejoras en el algoritmo de aprendizaje.

El trabajo con robots físicos impone ciertas restricciones a la hora de rentabilizar su aplicación en el campo industrial. La velocidad del algoritmo de aprendizaje presentado en esta tesis, hace necesaria la repetición de un buen número de ciclos de recogida para que los agentes determinen la estrategia de cooperación más adecuada para un ambiente. En el caso de robots reales, este tiempo de aprendizaje puede limitar su aplicabilidad. Una de las siguientes vías de trabajo que emprenderemos será la inclusión de distintas medidas de evaluación continua de la eficacia para incrementar la velocidad del aprendizaje. En este mismo sentido de mejoras en el algoritmo de aprendizaje, planeamos compaginar mecanismos similares a la *selección de parentesco* con medidas de la eficacia individual de los agentes.

Nuevos comportamientos sociales.

Otro objetivo inmediato que nos proponemos abordar es la ampliación del espectro de comportamientos sociales implementables en el colectivo de robots. Para ello, utilizaremos de nuevo los conocimientos de la biología para utilizar otras formas de interacción social presentes en la naturaleza. Estos estarán relacionados con los fenómenos de competición,

mutualismo, simbiosis, especialización etc. Con seguridad, la inclusión de estas formas de interacción entre agentes ampliarán la potencialidad y la eficacia de los sistemas multiagente.

Adaptación simultánea de varios comportamientos.

El proceso de adaptación utilizado en la tesis se enfoca sobre el aprendizaje de un único comportamiento cooperativo. Es de sumo interés conseguir mecanismos de adaptación que se enfrenten a varias presiones selectivas de forma simultánea. Nos proponemos desarrollar nuevos mecanismos de adaptación que, utilizando distintas señales de refuerzo y procesos de diferenciación somática como la maduración, permitan el aprendizaje en paralelo de varios comportamientos.

Referencias Bibliográficas

Ackley, D.H. y Littman (1991) Interactions between learning and evolution. En: *Artificial Life II: Proceedings Volumen of Santa Fe Conference.* (C.G. Laungton, J.D. Farmer, S. Rasmussen and C. Taylor, Eds.), Addison-Wesley.

Agre, P.E. y Chapman, D. (1987). Pengi: An implemantation of a Theory of Activity. *Proceedings of the AAAI'87.* Seatle. 268-272.

Alcock, J. (1993). *Animal Behavior: An Evolutionary Approach*, 5th Ed. Sinauer Associates, Sunderland.

Alexander, R.D. (1974). The evolution of social behavior. *Annual Review of Ecology and Systematics.* **5**. 325-383.

Aoki, S. (1983). A new Taiwanese species of *Colophina* (Homoptera: Aphidoidea) producing large soldiers. *Kontyu,* **51**, 282-288.

Arkin, R.C. (1989). Motor schema-based mobile robot navigation. *International Journal of Robotics Research*, **8**, 92-112.

Arkin, R.C. (1992). Cooperation without communication: Multi-agent schema based robot navigation. *Journal of Robotics Systems*, **9**, 351-364.

Axelrod, R. (1984). *La Evolución de la Cooperación.* Alianza Universidad. Madrid.

Axelrod, R. y Dion, D. (1988). The further evolution of cooperation. *Science.* **242**. 1385-1390.

Axelrod, R. y Hamilton, W. (1981). The evolution of cooperation. *Science*. **211**. 1390-1396.

Balch, T. y Arkin, R.C. (1994). Communication in reactive multiagent robotic systems. *Autonomous Robots*, **1**, 1-25.

Baldwin, J.M. (1896). A new factor in evolution. *American Naturalist*, **30**, 441-451

Barto, A.G., Sutton, R.S. y Watkins C.J.C.H. (1989). *Learning and Sequential Decision Making*. COINS Technical Report, **89-95**.

Barto, A.G., Sutton, R.S. y Brouwer, P.S. (1981). Associative search network: A reinforcement learning associative memory. *Biological Cybernetics*, **40**, 201-211.

Bateson, 1979. *Mind and Nature*. Flamingo.

Beckers, R., Holland, O.E. y Deneubourg, J.L. (1994). From local actions to global task: stigmergy and collective robotics. (En R.A. Brooks, y P. Maes (eds)), *Artificial Life IV: Proceedings of the Fourth International Workshop on the Synthesis*

Brooks, R.A. (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*. **2**. 14-23.

Brooks, R.A. (1990a). Elephant don't play chess. En *Designing Autonomous Agents*. P. Maes Ed. A Bradford Book. MIT Press. Cambridge.

Brooks, R.A. (1990b). *The Behavioral Language; User's Guide*. Technical Report AIM-1127, MIT Artificial Intelligence Lab.

Brooks, R.A. (1991a). Intelligence without representation. *Artificial Intelligence*, **47**, 139-160.

Brooks, R.A. (1991b). Intelligence without reason. *Proceedings of 12th International Joint Conference on Artificial Intelligence*, 569-595.

Clift, D.T., Harvey, I. y Husbands, P. (1993). Explorations in evolutionary robotics. *Adaptive Behavior*, **2**, 73-110.

Clutton-Brock, T.H. y Parker, G.A. (1995). Punishment in animal societies. *Nature*. **373**. 209-216.

Colmenares F. y Gómez, J.C. (1994). El desarrollo del comportamiento: aspectos funcionales y evolutivos. En J. Carranza (ed.), *Etología. Introducción a la ciencia del comportamiento,* 119-136. Cáceres: Publicaciones de la Universidad de Extremadura.

Colombetti M., Dorigo, M. y Borghi, G. (1996). Behavior Analysis and Training, a methodology for behavior engineering. *IEEE Transactions on System, Man and Cybernetics [B]*, **26**, 365-380.

Colorni, A., Dorigo, M. y Maniezzo, V. (1992). Distributed optimization by ant colonies. (En F. Varela y P. Bourgine (eds.)), *Toward a Practice of Autonomous Systems: Proc. of the First European Conference on Artificial Life*, 113-142, MIT Press,Cambridge.

Corbara, B., Drogoul, A., Fresneau, D. y Lalande, S. (1992). Simulating the sociogenesis process in ant colonies with MANTA. (En F. Varela y P. Bourgine (eds.)), *Toward a Practice of Autonomous Systems: Proc. of the First European Conference on Artificial Life*, 224-235, MIT Press, Cambridge.

Deneubourg,, J.L., Clip, P.L. y Camazine, S.S. (1994). Ants, buses and robots self-organization of transportation systems. *Proceedings of the Conference From Perception to Action*, 13-23, IEEE Computer Society Press.

Deneubourg, J.L., Goss, S., Franks, N., Sendova-Franks, A., Detrain, C. y Chretien, L. (1991). The dynamics of collective sorting robot-like ants and ant-like robots. (En J.A. Meyer y S.W. Wilson (eds.)), *From Animals To Animats: Proceedings of the First International Conference on Simulation of Adaptative behavior*, 356-363, MIT Press, Cambridge.

Deneubourgh, J.L., Theraulax, G. y Beckers, R. (1992). Swarm-made architectures. (En F. Varela y P. Bourgine (eds.)), *Toward a Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life*, 123-133, MIT Press, Cambridge.

Domjam, M. (1993). *The Principles of Learning and Behavior*. 3rd Ed. Brooks/Cole Publishing Co. Pacific Grove.

Dorigo, M. y Gambardella, L.M. (1995). ANT-Q: A reinforcement learning approach to combinatorial optimization. Technical Report No. **95-01**, IRIDIA, U.L.B.

Dorigo, M., Maniezzo, V. y Colorni, A. (1995). The Ant System: Optimization by a colony of cooperating agents. *IEEE Transactions on Systems, Man, and Cybernetics*.

Drogoul, A. y Ferber, J. (1993). From Tom Thumb to the dockers: Some experiments with foraging robots. En: J-A. Meyer, H.L. Roitblat y S.W. Wilson (eds.), *From animals to animats II: Proc. of the 2nd International Conference on Simulation of Adaptive Behavior*, 451-459, MIT Press, Cambridge.

Dugatkin, L.A. (1988). Do guppies play tit for tat during predator inspection visits?. *Behavioral Ecology and Sociobiology*. **23**. 395-399.

Dugatkin, L.A. (1991). Dynamics of the TIT FOR TAT strategy during predator inspection in the guppy (*Poecilia reticulata). Behavioral Ecology and Sociobiology.* **29,** 127-132.

Edelman, G. M. (1987). *The Neural Darwinism: The theory of Neuronal Group Selection.* New York, Basic Books.

Edelman, G.M., Reeke, G.N., Gall, W.E., Tononi, G., Williams, D. y Sporns, O. (1992). Synthetic neural modeling applied to a real-world artifact. *Proceedings of the National Academy of Sciences. USA*, **89**, 7267-7271.

Fletcher, D.J.C. y Michener, C.D. (1987). *Kin Recognition in Animals*. Wiley. New York.

Floreano, D. y Mondada, F. (1996). Evolution of plastic neurocontrollers for situated agents. En: *From Animals to Animats IV. (P. Maes, M. Mataric, J.A. Meyer, J. Pollack y S. Wilson Eds.). MIT Press, Cambridge.*

Foster, W.A. (1990). Experimental evidence for effective and altruistic colony defence against natural predators by soldiers of the gall-forming aphid *Pemphigus spyrothecae* (Hemiptera:Pemphigidae). *Behavioral Ecology and Sociobiology.* **27**. 421-430.

Goss, S. y Deneubourg, J.L. (1992). Harvesting by a group of robots. (En F. Varela y P. Bourgine (eds.)), *Toward a Practice of Autonomous Systems: Proc. of the First European Conference on Artificial Life*, 195-204, MIT Press,Cambridge.

Gullapalli, V. (1995). Skillfull control under uncertainty via direct reinforcement learning. *Robotics and Autonomous Systems*, **15**, 237-246.

Hamilton, W.D. (1964). The genetical theory of social behaviour. I-II. *Journal of Theoretical Biology*. **7**, 1-52.

Hammerstein, P. y Selten, R. (1993). Evolutionary game theory, En: *Handbook of Game Theory with Economic Applications* (Aumann R.J. & Hart S. eds.) Amsterdam. Holanda.

Harley, C.B. (1981). Learning the evolutionarily stable strategy. *Journal of Theoretical Biology*, **89**, 611-633.

Hashimoto, H., Takashi, K., Kudou, M. y Harashima, F. (1992). Self-organizing visual servo systems based on neural networks. *IEEE Control Systems*. April, 31-36.

Hecht-Nielsen, R. (1990). *Neurocomputing*. Addison-Wesley.

Herper, G. (1991). *Kin Recognition.* Cambridge University Press. Cambridge.

Hidalgo, S.J. (1994). Evolución de los comportamientos egoístas y cooperativos. En: *Etología: Introducción a la Ciencia del Comportamiento* (J. Carranza Ed.), Universidad de Extremadura, Cáceres, pp. 153-179.

Hinton, G.E. (1989). Connectionist learning procedures. *Artificial Intelligence,* **40**, 185-234.

Holland, J.H. (1992). *Adaptation in Natural and Artificial Systems.* MIT Press. Cambridge. MA.

Hölldobler, B. y Wilson, E.O. (1994). *Viaje a las Hormigas: una historia de exploración científica.* Ed. Grijalbo Mondadori.

Itô, Y. (1989). The evolutionary biology of sterile soldieres in aphids. *Trends in Ecology and Evolution.* **4**. 69-73.

Kaelbling, L.P. (1990). Learning in embedded systems. PhD Thesis. Stanford University.

Kaelbling, L.P., Littman, M.L. y Moore A.W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research,* **4**, 237-285.

Kazelnik, A. y Bernstein, C. (1994). Modelos de optimalidad en etología. En: *Etología: Introducción a la Ciencia del Comportamiento* (J. Carranza Ed.). Universidad de Extremadura. Cáceres, pp. 153-179.

Kohonen, T. (1988). *Self-organization and Associative Memory.* Second edition. Berlin: Springer-Verlag.

Koza, J.R. (1992). *Genetic Programming: On the Programming of Computers by Means of Natural Selection.* MIT Press. Cambridge. .

Krebs, J.R. y Davies, N.B. (1993), *An Introduction to Behavioral Ecology.* 3rd Ed. Blackwell Scientific Publications. Oxford.

Kube, C.R. y Zhang, H. (1993a). Collective robotics intelligence. En: J-A. Meyer, H.L. Roitblat y S.W. Wilson (eds.), *From animals to animats II: Proc. of the 2nd International Conference on Simulation of Adaptive Behavior,* 460-468, MIT Press, Cambridge.

Kube, C.R. y Zhang, H. (1993b). Collective robotics: From social insects to robots. *Adaptive Behavior,* **2**, 189-218.

Kube, C.R. y Zhang, H. (1994). Stagnation recovery behaviors for collective robotics. *Proceedings of the 1994 IEEE/RSJ International Conference on Intelligent Robots and Systems,* 1883-1890.

Lin, L.J. (1992). Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning,* **8**, 293-321.

MacLennan, B.J. y Burghardt, G.M. (1994). Synthetic ethology and the evolution of cooperative communication. *Adaptive Behavior*, **2**, 161-188.

Maes, P. (1991). Situated agents can have goals. En *Designing Autonomous Agentes*. (P. Maes Ed). A Bradford Book. MIT Press. Cambridge.

Mahadevan, S. y Connell, J. (1992). Automatic programming of behavior-based robots using reinforcement learning. *Artificial Intelligence*, **55**, 311-365.

Martín, P. y Millán, J. del R. (1997a). Learning reaching strategies through reinforcement for a sensor-based manipilator. *Neural Networks* (en prensa).

Martín, P. y Millán, J. del R. (1997b). A modular reinforcement based neural controller for a three-link manipulator. *IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Mataric, M. (1993). Designing Emergent Behaviors: From Local Interactions to Collective Intelligence. En J.-A. Meyer, H.L. Roitblat y S.W. Wilson (eds.), *From animals to animats II: Proceedings of the Second International Conference on Simulation of Adaptive Behavior*, 432-441. Cambridge, MA: MIT Press.

Mataric, M. (1994a). Reward functions for accelerated learning. En W.W. Cohen y H. Hirsh (eds.), *Machine Learning: Proceedings of the Eleventh International Conference*, 181-189. San Mateo, CA: Morgan Kaufmann.

Mataric, M. (1994b). Interaction and Intelligent Behavior. PhD Thesis. MIT.

Mataric, M. (1994c). Learning to behave socially. In D. Cliff, P. Husbands, J.-A. Meyer & S.W. Wilson (eds.), *From Animals to Animats: Third International Conference on Simulation of Adaptive Behavior*, pp. 453-462. Cambridge, MA: MIT Press.

Mataric, M. (1995). Designing and understanding adaptive group behavior. *Adaptative Behavior*, **4**, 51-80.

Mataric, M. (1996). Learning in multi-robots systems. En: Adaptation and learning in multi-agent systems. *Lecture notes in Artificial Intelligence*, 1041. Eds: Weiss,G., Sen, S. Berlin: Springer-Verlag, pp 152-163.

Mataric, M. (1997). Behavior-based control: examples from navigation, learning and group bahavior. *Journal of Experimental and Theoretical Artificial Intelligen-ce*, **9**, 323-336.

Maynard-Smith, J. (1964). Group selection and kin selection. *Nature*, **201**, 1145-1147.

Maynard-Smith, J. (1982). *Evolution and the Theory of Games*. Cambridge University Press.

Maynard-Smith, J. (1984). Game Theory and the evolution of behavior. *Behavioral Brain Science,* **7**, 95-125.

Maynard-Smith, J. y Price, G.R. (1973). The logic of animal conflict. *Nature.* **246**. 15-18.

McFarland, D. (1977). Decision making in animals, *Nature*, **269**, 15-21.

McFarland, D. (1993). *Animal Behavior.* 2nd Ed. Longman Scientific and Technical. Essex.

McFarland, D. (1994). Animal robotics: from self-sufficiency to autonomy. *Proceeding of the Conference From Perception to Action.* 47-54. IEEE Computer Society Press.

McFarland, D. y Boesser, T. (1993). *Intelligent Behavior in Animals and Robots.* MIT Press. Cambridge.

Meyer, J.A. y Guillot, A. (1990). *From Animals To Animats: Everything you Wanted to Know About the Simulation of Adaptive Behavior*, Technical Report BioInfo-90-1, Ecole Normale Superieure, París.

Milinski, M. (1987). TIT FOR TAT in sticklebacks and the evolution of cooperation. *Nature.* **325**. 433-435.

Milinski, M. y Parker, G. (1993). Competition for resources. En: (J.R. Krebs y N.B. Davies Eds.) *An Introduction to Behavioral Ecology.* 3rd Ed, Blackwell Scientific Publications, Oxford, pp.137-168.

Millán, J. del R. (1995). Reinforcement learning of goal-directed obstacle-avoiding reaction strategies in an autonomous mobile robot. *Robotics and Autonomous Systems*, **15**.

Millán, J. del R. (1996), Rapid, safe and incremental learning of navigation strategies, *IEEE Transactions on Systems, Man and Cybernetics [B]*, **26**, 408-420.

Millán, J. del R. y Torras, C. (1992). A reinforcement connectionist approach to robot path finding in non-maze-like environments. *Machine Learning*, **8**, 363-395.

Mitchell, T. M. y Thrun, S. B. (1993). Explanation-based neural networks learning for robot control. En C. L. Giles, S. J. Hanson y J. D. Cowan (eds.). *Advances in Neural Information Processing Systems 5*, 287-294. San Mateo, CA: Morgan Kaufmann.

Murciano, A. (1995). Aprendizaje de comportamientos cooperativos en sociedades de agentes autónomos. *Tesis Doctoral.* Universidad Complutense de Madrid.

Murciano, A. y Millán, J. del R. (1996). Learning signaling behaviors and specialization in cooperative agents. *Adaptative Behavior*, **5**, 5-29.

Murciano, A., Millán, J.del R. y Zamora, J. (1997). Specialization in multi-agent systems through learning. *Biological Cybernetics*, **76**, 375-382.

Murciano, A. y Zamora, J. (1993). Learning through adaptive value: a model working in a variable environment. En Gielen y Kappen (eds.), *Proceedings of the International Conference on Artificial Neural Networks*, 55-58. Springer-Verlag.

Murciano, A., Zamora, J. y Reviriego, M. (1993). A model for centering visual stimuli through adaptive value learning. En Mira, Cabestany y Prieto (eds.), *New Trends in Neural Computation*, 20-23. Springer-Verlag.

Murciano, A., Zamora, J., M. De la Paz, F., Girón, J.M. y Millán, J. del R. (1997). Robot móvil pra inestigación en grupos cooperantes, *XVIII Jornadas de Automática* CEA-IFAC, Gerona. 125-131.

Nolfi, S. (1997). Using emergent modularity to develop control systems for mobile robots. *Adaptive Behavior*, **5**, 343-363.

Numaoka, C. (1994). Phase transitions in instigated collective decision making. *Adaptive Behavior*, **3**, 185-223.

Numaoka, C. y Takeuchi, A. (1993). Collective choice of strategic type. En J.-A. Meyer, H.L. Roitblat y S.W. Wilson (eds.), *From animals to animats II: Proceedings of the Second International Conference on Simulation of Adaptive Behavior*, 469-477. Cambridge, MA: MIT Press.

Parker, L. (1994a). Alliance: An architecture for fault tolerant, cooperative control of heterogeneous mobile robots. *Proceedings of the 1994 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 776-783.

Parker, L. (1994b). Heterogeneuous Multi-Robots cooperation. PhD Thesis. MIT.

Pike, G.H., Pulliam, H.R. y Charnov, E.L. (1977). Optimal foraging: A selective review of theory and tests. *Quaterly Review of Biology*. **52**. 137-154.

Pitcher, T., Green, D. y Magurran, A.E. (1986). Dicing with death: predator inspection behavior. *Journal of Fish Biology*. **28**. 1439-1448.

Poulin, R. y Vickery, W.L. (1995). Cleaning symbiosis as an evolutionary game: to cheat or not to cheat?. *Journal of Theoretical Biology*. **175**. 63-70.

Prescott, T.J. y Mayhew, J.E.W. (1992). Obstacle avoidance through reinforcement learning. In J.E. Moody, S.J. Hanson, y R.P. Lippmann (eds.), *Advances in Neural Information Processing Systems 4*, 523-530. San Mateo, CA: Morgan Kaufmann.

Rapoport, A. y Chammah, A.M. (1965). *Prisoner's Dilemma.* University of Michigan Press. Ann Harbor.

Reeke, G.N., Sporns, O. y Edelman, G.E. (1990). Synthetic neural modelling: The "Darwin" series of recognition automata. *Proceedings IEEE.* **78**, 1498-1530.

Reeke, G.N., Finkel, L.H., Sporns, O. y Edelman, G.M. (1990). En: Signal and Sense: Local and global order in perceptual maps. Eds: Edelman G. M., Gall, W.E. y Cowan, W.M. (Willey, New York), pp 607-707.

Reyer, H.U. (1984). Investement and relatedness. A cost/benefit analysis of breeding and helping in the pied kingfisher (*Ceryle rudis*). *Animal Behavior.* **32**, 1163-1178.

Reyer, H.U. (1986). Breeder-helper interactions in the pied kingfisher reflect the costs and benefits of cooperative breeding. *Behavior* **96**. 277-303.

Rumelhart, D.E., Hinton, G.E. y Williams, R.J. (1986). Learning representations by back-propagating errors. *Letters to Nature*, **323**, 533-535.

Steels, L. (1994a). A case study in the behavior-oriented design of autonomous agents. En D. Cliff, P. Husbands, J.-A. Meyer y S.W. Wilson (eds.), *From Animals to Animats III: Third International Conference on Simulation of Adaptive Behavior*, 445-452. Cambridge, MA: MIT Press.

Steels, L. (1994b). The artificial life roots of artificial intelligence. *Journal of Artificial Life*, **1**, 89-125.

Steels, L. (1994c). Emergent functionality in robotic agents through on-line evolution. En R.A. Brooks y P. Maes (eds.), *Artificial Life IV: Proceedings of the Fourth International Workshop on the Synthesis and Simulation of Living Systems.* 8-14. Cambridge, MA: MIT Press.

Steels, L. (1996). Discovering the competitors. *Adaptive Behavior* , **4**, 173-199.

Stephens, D.W. y Krebs, J.R. (1986). *Foraging Theory.* Princeton University Press. Princeton.

Stephens, D.W., Nishimura, K. y Toyer, K.B. (1995). Error and discounting in the iterated Prisoner's dilemma. *Journal of Theoretical Biology*, 176, 457-469.

Sutton, R.S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. *Proc. of the 7th International Conference on Machine Learning*, 216-224.

Sutton, R.S. y Barto, A.G. (1987). A temporal-difference model of classical conditioning. *Proceedings of the Ninth Annual Conference of the Cognitive Science Society.* Seattle, WA.

Tan, M. (1993). Multi-agent reinforcemente learning: Independent vs. cooperative agents. *Proceedings of the 10th International Conference on Machine Learning*, 330-337.

Theraulaz, G., Goss, S., Gervet, J. y Deneubourg, J.L. (1991). (En J.A. Meyer y S.W. Wilson (eds.)), *From Animals To Animats: Proceedings of the First International Conference on Simulation of Adaptative behavior*, 356-363, MIT Press, Cambridge.

Thuijsman, F., Peleg, B. Amitai, M. y Shmida, A. (1995). Automata, matching and foraging behavior of bees. *Journal of Theoretical Biology*. **175**. 305-316.

Torras, C. (1985). *Temporal-pattern learning in neural models*. Lecture Notes in Biomathematics No. 63. Springer-Verlag.

Trivers, R.L. (1971). The evolution of reciprocal altruism. *Quaterly Review of Biology*. **46**.35-57.

Trivers, R.L. (1985). *Social Evolution*. Benjamin Cummings, Menlo Park, California.

Wagner, G.P. y Altenberg. L. 1996. Complex Adaptations and the evolution of evolvability, *Evolution,* **50,** 967-976.

Weiss, G. (1995). Distributed reinforcement learning. *Robotics and Autonomous Systems,* **15**:135-142

Werner, G.M. y Dyer, M.G. (1993). Evolution of herding behavior in artificial animals. En: J-A. Meyer, H.L. Roitblat y S.W. Wilson (eds.), *From animals to animats II: Proc. of the 2nd International Conference on Simulation of Adaptive Behavior*, 502-510, MIT Press, Cambridge.

Wilkinson, G.S. (1984). Reciprocal food sharing in the vampire bat. *Nature*. **308**. 181-184.

Williams, G.C. (1966). *Adaptation and Natural Selection*. Princeton Univ. Press. Princeton.

Wilson, D.S. (1979). *Natural Selection of Populations and Communities*. Benjamin Cummings, Menlo Park.

Wilson, E.O. (1971). *The Insects Societies*. Harvard University Press. Cambridge.

Wilson, E.O. (1975). *Sociobiology: The New Synthesis*. Belknap Press, Cambridge.

Wilson, S.W. (1985). Knowledge Growth in an Artificial Animal. En: *Proceedings of the First International Conference on Genetic Algorithms and their Applications*, (J.Greffensette Ed), Lawrence Erlbaum Associates.

Wynne-Edwards, V.C. (1962). *Animal Dispersion in Relation to Social Behavior*. Oliver & Boyd. Edimburgo.

Wynne-Edwards, V.C. (1986). *Evolution through Group Selection*. Blackwell Scientific Publications, Oxford.

Zamora. J., Millán, J. del R. y Murciano, A. (1997). Learning and stabilization of altrustic behaviors in multiagent systems. *IEEE International Symposium on Computational Intelligence in Robotics and Automation,* IEEE Computes Society Press, pp. 287-293.