Croatian Journal of Philosophy

Vol. XXII · No. 66 · 2022

Croatian Journal of Philosophy

1333-1108 (Print)
1847-6139 (Online)

Editor:
Nenad Miščević (University of Maribor)

Advisory Editor:
Dunja Jutronić (University of Maribor)

Managing Editor:
Tvrtko Jolić (Institute of Philosophy, Zagreb)

Editorial board:
Stipe Kutleša (Institute of Philosophy, Zagreb),
Davor Pećnjak (Institute of Philosophy, Zagreb)
Joško Žanić (University of Zadar)

Advisory Board:
Elvio Baccarini (University of Rijeka), Carla Bagnoli (University of Modena), Boran Berčić (University of Rijeka), István M. Bodnár (Central European University), Vanda Božičević (Bergen Community College), Sergio Cremaschi (Milano), Michael Devitt (The City University of New York), Peter Gärdenfors (Lund University), János Kis (Central European University), Friderik Klampfer (University of Maribor), Željko Loparić (Sao Paolo), Miomir Matulović (University of Rijeka), Snježana Prijić-Samaržija (University of Rijeka), Igor Primorac (Melbourne), Howard Robinson (Central Europena University), Nenad Smokrović (University of Rijeka), Danilo Šuster (University of Maribor)

Co-published by
"Kruzak d.o.o."
Naserov trg 6, 10020 Zagreb, Croatia
fax: + 385 1 65 90 416, e-mail: kruzak@kruzak.hr
www.kruzak.hr

and

Institute of Philosophy
Ulica grada Vukovara 54/IV, 10000 Zagreb, Croatia
fax: + 385 1 61 50 338, e-mail: filozof@ifzg.hr
www.ifzg.hr

Available online at http://www.ceeol.com and www.pdcnet.org

CROATIAN JOURNAL OF PHILOSOPHY

Vol. XXII · No. 66 · 2022

Kathleen Vaughan Wilkes (1946–2003)

Book Review

Introduction

Inter-University Centre Dubrovnik (IUC), St Hilda's College Oxford and the Herbert Simon Society in Milan run the annual Kathy Wilkes Memorial Conference *collaboratively. The three institutions will be taking it in turns to host the event, starting with Dubrovnik in 2022. In 2023 it will be held in Oxford and in 2024 it will be held in Milan and Turin.*

The papers published in this issue of the Croatian Journal of Philosophy *were presented at the first event that took place in Dubrovnik in April 2022. The guest editor of this issue feels that there is no need for any introductory remarks about Kathy Wilkes since the first two papers by Nada Bruer and Paul Fletcher give factual information together with fond personal memories of Kathy Wilkes as a philosopher, defender of academic freedom, friend to many and much much more. The rest of the papers deal either with her work or discuss isssues directly or indirectly inspired by her many ideas.*

DUNJA JUTRONIĆ

Croatian Journal of Philosophy
Vol. XXII, No. 66, 2022
https://doi.org/10.52685/cjp.22.66.1
Received: July 5, 2022
Accepted: August 16, 2022

Kathy Wilkes at the Inter-University Centre Dubrovnik: Philosophy, Courage, and much more

NADA BRUER LJUBIŠIĆ
Inter-University Centre, Dubrovnik, Croatia

The text presents the activities of Dr. Kathleen Vaughan Wilkes, a philosopher from the University of Oxford in the Inter-University Centre Dubrovnik (IUC) from the beginning of the 1980s to the end of the millennium. Dr. Wilkes was co-directing the longest standing IUC course Philosophy of Science*, but she also initiated other IUC academic programmes. As a member of the IUC governing bodies, she was highly engaged in securing scholarships for participants from Central and East Europe in IUC programmes, mostly through Open Society Foundation. Dr. Wilkes played a crucial role in spreading information from the city of Dubrovnik during the attacks of the Yugoslav People's Army in 1991 and during Croatian's struggle for independence, for which she was awarded honorary citizenship and posthumously one of the squares was named after her.*

Keywords: Kathy Wilkes; Inter-University Centre (IUC); Philosophy of Science; Open Society Foundation; Central and East-European scholars; dissemination of information.

It is no wonder that Kathy Wilkes found her way to the Inter-University Centre Dubrovnik (IUC).[1] Dr. Kathleen Vaughan Wilkes, a Tutor and Fellow in Philosophy at Oxford's St. Hilda's College was passionate not only about philosophy but also about social changes. She was initiated to the IUC by her Oxford colleague Dr. William Newton-Smith, who invited her to take part in the IUC *Philosophy of Science* course in

[1] In the text, Dr. Wilkes would mostly be referred with the less formal version of the name—as Kathy, as her friends, colleagues and later on citizens of Dubrovnik called her.

April of 1981. The theme of the course that year was "Theories and Explanations." Kathy Wilkes came as a resource person and held a lecture on reductionism. From that year on, she never missed this April course, which she had already started co-directing in 1984.[2]

In the early 80s, the IUC operated for almost ten years as an independent international institution for advanced studies. It was envisioned and launched between 1970 and 1972 by Prof. Ivan Supek, at that time Rector of the University of Zagreb, to offer the academic community a free platform to develop international cooperation independent of governmental control and national constraints. Supek's idea was that by "preparing the ground for the present day scientific revolution, the university community has also prepared the ground for a better world, the world of human understanding and peace" and to address urgent world problems, new organisational structures needed to be offered (Supek 1971: 1). Dubrovnik was chosen as a seat of this new institution so that its centuries-long history of an independent city republic between East and West, North and South, using diplomacy to secure freedom and economic stability, would inspire contemporary scientists to explore current social developments. Another political fact contributed to the convenient location of the Centre in the world divided by an Iron Curtain. Former Yugoslavia, although a communist state, was among the leaders of the Non-aligned Movement which meant that people from Eastern and Western blocks could come there and the IUC became a meeting place of scholars along the division line. As Prof. James Robert Brown, Dr. Wilkes's colleague and co-director of the *Philosophy of Science* course wrote, Dubrovnik "was liberal-minded and cosmopolitan, and it easily accommodated the tensions of the cold war" (Brown 2010: 36). The IUC and Dubrovnik soon became Kathy Wilkes's second home.

1. *Academic activities*

Philosophy of Science, also initiated by Ivan Supek, is one of the first programmes that started after the establishment of the IUC. The first course took place at the end of 1974 under the title *Philosophy of Science and Humanism: Foundations of Science and Theory of Knowledge* and one of the lecturers was the Nobel prize laureate Werner Heisenberg. From the very beginning participants in these courses were coming from Western European countries, USA and Canada, countries within the Eastern block and of course from former Yugoslavia. The first courses lasted almost a month, but due to organisational issues, they were held for two weeks, and later on, only one. When Kathy Wilkes started co-directing it in 1984, her colleague organisers were Lars Bergström from the University of Uppsala, Wladyslaw Krajewski from the University of Warsaw, Srđan Lelas from the University of Zagreb,

[2] Data from the Inter-University Centre Dubrovnik archive.

Jürgen Mittelstrass from University of Konstanz, William Newton-Smith and Rom Harré from Oxford University. In 1984 William Newton-Smith and Kathy Wilkes edited papers deriving from the course in the journal *Ratio,* to be published the following year. The success of this publication led them to establish the journal *International Studies in the Philosophy of Science*: "to promote the discussion of issues in the philosophy of science by those of differing philosophical, cultural and political backgrounds" (Newton-Smith and Wilkes 1986: 2). The journal was planned to be biannual and in these two issues, it was to cover themes from the course. The first edition emerged in 1986 and contained historical studies covering the period from Galileo to Newton. Editors attempted to search for a wide range of countries to ensure the international exchange of ideas. Soon, the journal started including book reviews as well. From 1990 on, the journal had three annual issues, while in 1992, the editorial board was joined by dr. Riccardo Viale from Istituto di Methodologia della Scientia e della Technologia from Torino. After ten years of continuous work, the editorial position of the journal has been taken over by another co-director of the *Philosophy of Science* course, Prof. James Robert Brown, from University of Toronto.[3]

Besides extensive work on *Philosophy of Science*, Kathy Wilkes also co-directed other programmes within the IUC. Already in March 1983, along with David Charles, Timothy Williamson, Aleksandar Pavković, David Brown, and Neven Sesardić, she co-organised the course *Contemporary Issues in the Philosophy of Mind: Functionalism and Explanation* where she held a lecture on "Varieties of Functionalism." The following year in fall the same group of scholars organised the programme *Truth and Knowledge.* Soon, Živan Lazović and Miloš Arsenijević from the University of Belgrade also joined this group and the programme was held until the beginning of the 1990s. Also, in the mid 1980s, together with William Newton-Smith and other colleagues from the region, under the auspices of and sponsored by the Soros Foundation from New York, she helped organise programmes *The Culture of Central and Eastern Europe* that soon became the programme *Central and Eastern Europe in Transformation.* These course series included topics on historical events and fiction, transition from dictatorship to freedom, economic sociology in comparative perspectives, political technology of reforms, the cultural history of central Europe and many others.[4] The

[3] The current editorial board of the *International Studies in the Philosophy of Science* is still encouraging participation in the annual *Philosophy of Science* conference at the IUC and publishing accepted papers deriving from this programme. From 2010 it consists of 4 yearly issues.

[4] Steering committee of this programme through different years consisted of Pavel Cmorej from Bratislava, Ladislav Hojdánek and Jan Havranek from Prague, Imre Hronszky and Tibor Vamos from Budapest, Dejan Kjuranov from Sofia, Carl E. Levitin, Yuri Afanasiev and Boris Raushenbakh from Moscow and Andrzej Ziabickiand, Klemens Szaniawski and Wlodzimierz Siwinski from Warsaw, and Ante Stamać from Zagreb.

programme was in 1990 held, but by 1991, despite announcements, former Yugoslavia was transforming in a way that war activities prevented the regular implementation of IUC events.

But one of the features of the IUC programmes, as participants always testify and work schedules also confirm, is that additional inspiration for the academic discussion, for the exchange of knowledge and ideas is received from gatherings outside the classroom, in the buildings' courtyard, in nearby restaurants, beaches or during different social or extracurricular activities.[5] Being a warm and approachable person, Kathy was equally fond of these social gatherings that helped her create a strong network of colleagues and friends.

2. *Member of the IUC Executive Committee to secure grants*

At the 8th Meeting of the IUC Council, in August 1985, Kathy Wilkes was elected to the IUC governing bodies as a member of the Executive Committee. Right away she widened her engagement beyond her philosophical courses. One of her main activities was ensuring that scholars from the Eastern bloc countries could come and participate in the IUC international programmes, a task that she easily took upon herself but that was far from easy.

When Kathy Wilkes started coming to the IUC Dubrovnik, she had a rich experience cooperating with dissident philosophers in Central and East Europe. In 1979, she accepted an invitation from Czech philosopher Julius Tomin and traveled to Prague to hold informal philosophical seminars organised in his home. These meetings were under the surveillance of the police and participants were often intimidated. Kathy Wilkes was not to be intimidated. She made three trips to Prague until she was denied a visa. However, she continued supporting philosophers behind the Iron curtain. Through the Jan Hus Educational Foundation, she helped to organise other colleagues from Oxford to travel to Prague and kept sending books. Later, she started going to other East European countries: Poland and Bulgaria. So, at the IUC, Kathy just continued her mission.

Kathy Wilkes and William Newton-Smith had a strong connection with George Soros and his foundation, which was opened in Hungary in 1984 but later also in other countries with an incentive to help spread information in Eastern European countries and Russia. This led to the establishment of the Open Society foundations network. Through them, Wilkes and Newton-Smith secured funds for the participation in IUC programmes. The fund was called "Grants for younger scholars"

[5] For example, a football match "Yugoslavia against the rest of the World" was organised for the course *Contemporary Issues in the Philosophy of Mind* with the score 1:1, while for example Polish/Russian vodka party was organized for the *Philosophy of Science* course in April 1988.

(OSF support). It was intended for scholars under 40 years of age from the countries of Central and East Europe. The scholarship covered full board and accommodation in Dubrovnik during the course, excluding travel. Each year the IUC announced the list of eligible courses for the OSF support, which were in the humanities and social sciences. Courses in medicine or natural sciences were not financially supported.

The IUC archive is full of documents from the second half of the 80s of Kathy's engaged correspondence with programme coordinators of the Soros Foundation in Budapest, discussing different aspects of the scheme and eligibility criteria.[6] Occasionally, she contacted George Soros as well advocating different solutions. The grants were sometimes difficult to administer since participants from different countries applied either to foundations in their local countries—as soon as they were established (SSSR, Poland, Hungary)—or directly to Kathy Wilkes (Albania, Bulgaria, Czechoslovakia, East Germany, Romania, Yugoslavia). The "Younger Scholars" grant scheme was highly successful. From 37 grants administered in the year 1987/88 to 233 grants for the year 1989/90 with approximately 1500 applications. In her reports, Kathy Wilkes always paid attention to the distribution of grants according to countries urging the IUC to encourage universities from the under-represented areas to join, in order to make this possibility available to their scholars as well.

But besides the cooperation with OSF and administering this scholarship, Kathy Wilkes additionally initiated and secured the support to young Eastern European scholars from the New York Foundation. The IUC received 5000 USD in 1986 and few consecutive years from this source. Kathy just did everything in her power to find a way to bring people together. The longstanding Executive Secretary of the IUC Ms. Berta Dragičević testified that Kathy was "immensely brave. Since the IUC could not have an account for foreign currency, she would bring money in her purse. Then, she would divide it to people from East and Middle Europe, later from China. Kathy was a cosmopolitan who wholeheartedly worked on connecting philosophers from East and West" (Rudež 2017: 56–57).

In 1990 all East European countries were undergoing political and social changes and the Open Society Foundation/IUC scheme administered through Oxford was no longer in operation. All former communist countries started having their home foundations and in her report to the IUC Executive Committee in October 1990, Kathy testified about visiting Bulgarian, Czechoslovak and Romanian Foundations and shared their determination to set aside funds for their citizens taking part in the IUC programmes. Kathy Wilkes concluded in her report: "I am relieved that the 'OSF/IUC' scheme is no longer needed. It pro-

[6] She argued that the age limit for grant applicants should be raised from 35 to 40 or 45, attempted to secure more funds and paid attention that the IUC course fees were covered as well.

duced splendid results, and was necessary at the time, but its popularity led to an inordinate amount of work" (Wilkes 1990). Yugoslav foundation was to be opened in June 1991, but by then, the country was to fall apart in military aggression and war.[7] Slowly, IUC courses were to be canceled. But not all. *The Philosophy of Science* conference continued to meet, even during the siege of Dubrovnik.

3. *Attacks on Dubrovnik*

At the end of the summer of 1991 Dubrovnik and its surroundings (as other parts of Croatia even before) started being attacked by Yugoslav People's Army, by then deprived of Slovenian and Croatian leadership since these countries had already held referendums and declared their independence from former Yugoslavia. These events caught Kathy Wilkes on site as the Chair of the Executive Committee (EC), the position she was elected to in August 1988. She decided not to return to her home in Oxford, but to stay in Dubrovnik since suddenly there was new work to be done. The situation became graver and graver. Bombs from the air, land and sea were falling on the city and its surroundings, the city was under blockade, without running water or electricity, and hotels that accommodated tourists until recently accepted refugees from the neighboring villages. While communication channels at the IUC were in function, she was working with the IUC staff from the IUC office, sending out faxes and letters and making phone calls daily. Soon, as the telephone and fax lines became functioning less, on the request of the Dubrovnik mayor, Mr. Pero Poljanić, she and Berta Dragičević moved to the fort of St. John and later to the municipality office and used one of the three satellite connections in the city to continue sharing information from Dubrovnik (Dragičević 2017: 75–82). As soon as the direct bombing of the city started, Kathy contacted the BBC and reported on the situation, the task she would continue almost daily during the city's siege. She also worked as a personal assistant to the Mayor Poljanić, translating his numerous appeals that she later learned had reached both Margaret Thatcher and Prince Charles (Dedo 2021: 10–11). Kathy also wrote her letters to different magazines: *The Times*, *Independent, Guardian*, *Observer*, *Sunday Times*, *Sunday Telegraph*, *Financial Times*, and *New York Times*. With the same energy she had to establish connections along the Iron Curtain, she now used her contacts in the British and world politics to spread a word of what was really going on in Dubrovnik and in Croatia. She approached President Vaclav Havel, whom she knew from their dissident days in Prague, George Soros, Lord Carrington (a family relation), Sir Alec Douglas Hume and many others, asking them to use their influence to stop the war in Croatia (Dragičević 2017: 75–82). As a foreigner and a distinguished Oxford philosopher, her testimonies were considered trustworthy, they helped in sharing the truth.

[7] OSF Croatia was established in 1992.

Besides writing to the world, Kathy also started writing to citizens of Dubrovnik; after all, she became one of them in the city under the blockade. From the 8th of November 1991 until the 6th of January 1992, a daily bulletin was printed in the city called *The Voice from Dubrovnik* (*Glas iz Dubrovnika).* It was a simple edition of news and texts printed in Croatian and English, distributed to Dubrovnik citizens and sent abroad to spread the information from the city. For 67 editions, Kathy Wilkes wrote 28 different texts commenting current situation, informing about letters sent from the city to the world by the mayor or herself or just observing and analysing the current situation and reflecting on human nature or the nature of the war. She was always oriented toward the future. Even very early on, in mid-November 1991, she reminded her fellow citizens that new kind of courage would be needed to rebuild relationships because "whatever the atrocities perpetrated upon Croatia in general and Dubrovnik in particular, it is impossible to pick up the country and move it elsewhere…it will take every last drop of the internationalism for which Dubrovnik has for so long been famous, to repair relationships there. But, somehow, it will have to be done" (Wilkes 1991a). And she never stopped being analytical. In another letter, she referred to rumours advising, "let us be more sceptical, pausing and checking before jumping to conclusions and then passing them on. And that goes for the newsmen too" (Wilkes 1991b).

Her activities during the war did not stop behind the typing machine and telephone sets. With Dr. Slobodan Lang on the 6th of November 1991, she co-organised the IUC conference on the *Quality of Life and Human Rights of Refugees in Dubrovnik.* She also left the city under siege to arrive in Mokošica, the occupied suburb area of Dubrovnik and spent a day with the Red Cross wanting to find out what kind of humanitarian aid to send.[8] On the 6th of December, the IUC building was hit by incendiary shells and it burnt down. Ten days later, Kathy Wilkes and Berta Dragičević sent letters to IUC friends and colleagues describing the destruction and asking for help with words: "For we had a dream in 1971: the project of uniting the world in Dubrovnik...This dream is now in ashes, not metaphorically but literally. We now have to start dreaming again; … we trust that we will have as many of you as possible sharing this new dream" (Øyen and Dragičević 2002: 15).

Kathy Wilkes continued working on the dream. She would leave the city only briefly to organise help.[9] At the beginning of January 1992, she left Dubrovnik, where the situation was improving, to return to her teaching position in Oxford. But again, she acted in different directions. She organised visits of Dubrovnik mayor to the University and the city of Oxford. As a Chair of the IUC EC, she took part in meetings in Vienna, Hamburg, Santa Barbara, Oxford and Zagreb to ensure the

[8] Based on the memory of Vesna Gamulin in Obradović Mojaš (forthcoming).

[9] One such trip was to the USA that lasted nine days. It took her three days to go there and three to come back.

continuation of the Inter-University Centre as an institution and to support the rebuilding process.[10] She continued gathering humanitarian aid and drove it in the truck[11] to Croatia. She also worked to secure mine removal equipment, being aware that the best way to help is to re-establish everyday life.

To support the academic activity at the IUC, she invited her colleagues to participate in the annual *Philosophy of Science* course, which was to take place in the Music school. It was one of the most memorable events for the IUC participants. Many came to work in the city that was occasionally bombed and testified that they were inspired by their philosophical discussions with music practice from the neighbouring rooms. Kathy Wilkes continued working for the IUC in the years after the war. She was the Chair of the EC until 1996 but continued coming to *Philosophy of Science* until 2003.

4. *Recognitions*

For all she has done for Dubrovnik, Kathleen Vaughan Wilkes was awarded honorary citizenship of Dubrovnik on the day of St. Blaise, the patron of the city, in February 1993 and the portrait of her is permanently placed in the City's Council Hall. Honouring not only her scientific work at the IUC and beyond but also the tremendous help in Croatia's struggle for independence and support for the development of the Croatian academic community, Dr. Wilkes was awarded Doctor Honoris Causa in the field of philosophy by the University of Zagreb on the 22nd of May, 2001.

Unfortunately, as her personal well-being was never her priority, her health was already deteriorating. James Robert Brown wrote that all that she has done "took a great emotional toll on her, but she would not have wanted it any other way" (Brown 2010: 37). She passed away in Oxford on the 21st of August 2003. Following her wishes, her ashes were scattered in the sea, in Dubrovnik Pile area, below the Fort Lovrjenac, which above the entrance has a Latin reminder from the time of the Dubrovnik Republic that says: "Liberty is not to be sold for any gold." With this saying, Kathy was in complete accord. Following her death and upon the initiative of Berta Dragičević in 2011 a small square in Pile area, in front of the church of St. George, with the view to the very same fort was named after her. The ceremony of presenting the memorial plaque organized by the city was held on the 1st of February, 2012.

[10] The first phase of the reconstruction of the IUC building, organised by the University of Zagreb and financed by the Croatian government was over in summer 1993.

[11] Prof. James Brown thinks that this might have been an ambulance which she bought in the UK.

5. *Legacy*

The work of Kathy Wilkes was never forgotten among the IUC participants, members of the governing bodies and friends. On many occasions anecdotes are told of her brave undertakings and her vibrant spirit. Nenad Miščević, Kathy's colleague from the *Philosophy of Science* wrote down "So, what do you say about the character of a physicalist guardian angel like Kathy, who would be offended if told she had a good soul or a great mind? Well, that she had a great heart. That's physicalistic enough" (Miščević 2010: 86).

During her life, together with William Newton-Smith, she established an Inter-University Foundation to be used to fund participants from former communist countries in the *Philosophy of Science* courses. Since the majority of these countries have eventually joined the EU and the fund was no longer used for that purpose, William Newton-Smith in 2011 secured that the remaining amount would be used in the memory of Kathleen Vaughan Wilkes to refurbish one of the classrooms in the IUC building, still unfinished after the war destruction. The room dedicated to Kathy Wilkes at the IUC now accommodates new generations of students and professors for their vivid discussions.

On April 16–17 2018 Dr. Anita Avramides and Dr. Paul Flather organised the Kathy Wilkes Memorial Conference at St Hilda's College to celebrate Kathy's memory and reflect on her important contributions to philosophy and politics. At this event, inspired by the speaker's enthusiastic reminiscences of Kathy's work, the idea came to establish a new conference series that would take place in her honour. It was then that St. Hilda College, Herbert Simon Society from Turin and Inter-University Centre Dubrovnik initiated a project of hosting conferences on Mind, Philosophy, and Society in the memory of Kathy Wilkes alternately in these three locations. The first conference on the topic of *Re(assessing) Goal-Directed Activity* took place in April 29–30, 2022 at the IUC in Dubrovnik and this volume is derived from that conference.

Kathy Wilkes, as a philosopher, tutor, and intellectual, but utmost as an extraordinary human being, is still inspiring people who were privileged to work with her, know her, or to have learned about her rich life.

References

Brown, J. R. 2010. "Dubrovnik Reminiscences." In Ø. Øyen and B. Dragičević (eds.). *Fragments of Memories of Life and Work at Inter-University Centre Dubrovnik 1971 – 2007.* Dubrovnik: IUC.

Dedo, I. 2021. "Informacijama nismo zaustavili rat, ali smo ljudima otvorili oči." *Du List* 29. September 2021: 10–11.

Dragičević, B. 2017. "Dissemination of Information as a Contribution to the City Defence." In R. de la Brosse and M. Brautović (eds.). *Reporting the Attacks on Dubrovnik in 1991, and the Recognition of Croatia.* Cambridge Scholars Publishing, 75–82.

Miščević, N. 2010. "My Three Decades at the IUC: A Testimony". In Ø. Øyen and B. Dragičević (eds.). *Fragments of Memories of Life and Work at Inter-University Centre Dubrovnik 1971 – 2007.* Dubrovnik: IUC, 83–88.

Newton-Smith, W. and K. Wilkes. 1986. "Introduction." *International Studies in the Philosophy of Science* 1 (1): 2.

Obradović Mojaš, J. Forthcoming. "Interuniverzitetski centar Dubrovnik u ogledalu Grada. Odabrane slike polustoljetnog postojanja." *Dubrovnik.*

Øyen, Ø. and B. Dragičević. 2002. *Beyond Frontiers. 30th Anniversary of Inter-University Centre Dubrovnik.* Dubrovnik: IUC.

Rudež, T. 2018. "Kathy Wilkes." *Jutarnji list,* the 21st of May 2018: 56–57.

Supek, I. 1971. *The Inter-University Centre for the Humanities and Social Sciences (IUCHSS) in Dubrovnik*. Zagreb. University of Zagreb.

Wilkes, K. W. 1990. *"OSF/IUC SCHEME 1989–90: final report for 45th EC Meeting*" – IUC archive documentation (October).

Wilkes, K. V. 1991a. *The Voice of Dubrovnik*, No. 11, the 18th of November.

Wilkes, K. V. 1991b. *The Voice of Dubrovnik*. No. 20, the 27th of November.

Croatian Journal of Philosophy
Vol. XXII, No. 66, 2022
https://doi.org/10.52685/cjp.22.66.2
Received: August 30, 2022
Accepted: October 2, 2022

Memories of Dubrovnik's Global Citizen—Kathy Wilkes

PAUL FLATHER
University of Oxford, Oxford, UK

This is a personal memoir about the life, work and courage of Professor Kathleen Wilkes, a Fellow in Philosophy for 30 years at St Hilda's College, Oxford University. The article traces—and sets out to explain—particularly her links to Dubrovnik and Croatia and the Inter–University Centre since 1981, and supported strongly through the 1980s and even during the 1990s, remaining on site during the cruel siege of the city when the IUC suffered a devastating fire. Three key aspects of her life are explored—her work as a significant philosopher of science; her outstanding courage and work in defending academic freedom widely over the East Central European region, and her warm personality and generous friendship. This is why she can be regarded as Dubrovnik's Global Citizen, the IUC was only too ready and willing to host this conference in her honour.

Keywords: Kathy Wilkes; Inter-University Centre; Dubrovnik; Oxford; St Hilda's College

This volume of papers from a recent conference at the Inter-University Centre in Dubrovnik is now published in fond memory of the philosopher, defender of academic freedom, and friend to many, Kathy Vaughan Wilkes.

It may seem a stretch to appreciate why a distinguished Fellow and Tutor in Philosophy at St Hilda's College, Oxford University, for 30 years, should be so honoured at a conference in the Philosophy of Science, held at the IUC in Croatia. But to all who knew her—and that was a privilege I shared—knew her for her devotion, solidarity and courage relating to the infamous 1991 siege of this evocative treasured medieval city. As such, this honour is entirely appropriate.

We, happy band of philosophers, fellow thinkers, friends, and supporters, gathered for this event with some pride in the main hall of the IUC, which Kathy so loved, in April 2022, to discuss a range of philosophical issues, including intentionality, goal-directed behaviour and behaviour regulation, machine learning, and action understanding, all of which would have had Kathy jumping in to discuss with many perspectives. Oh, if only she could have been present with us, too, brandishing her wit, flashing her dry humour, and keeping us to the mark, until, that it was time, for her to lead us all to enjoy and relish the "down time", when discussion would, of course, continue.

Before our proceedings were underway, we were treated to a tour of the famous city taking in certain key spots which had meant much to Kathy. We visited the small, picturesque Pile harbour, just before the main Pile gate entrance to the city, with its blue sea and undulating waves, where Kathy would sit and look out, so inviting for a dip, and, no doubt, sometimes she would swim there. But it was also here during the darkest days of the 1991 siege of Dubrovnik by the invading Serbs, with electricity cut off, with water and supplies extremely short, we were told that she once saw a dead body floating in those waters, and, by her own admission, the full horror of the war was brought home to her, savagely.

Back in 1991, having already spent a decade associated with the IUC which was now caught up in the crossfire of the siege as part of the War of Croatian Independence, Kathy, caught in the country at the time, instantly committed herself to supporting and defending the centre. She became unofficial English language assistant to the Mayor. She chose to spend long periods in the city, even taking a year of absence from her college teaching duties at Oxford, as we learnt from her close colleague, Professor Anita Avramides, to express personal solidarity through her presence. She even took to dressing in fatigues as befitted her sense of embattlement and would start off morning gatherings at the centre with a review of the latest battle lines to identify just where invading troops had reached. No wonder she was even made an Honorary Private in the Croatian army.[1]

In a memorable moment, during a lunchtime interview on a BBC radio news programme, which I happened to hear live, she was asked, as she sat in the IUC Library room, just how bad the situation was, and, indeed, to confirm what were the various background noises that could be heard, on air. "Oh, well, it must be firing and bombing," she replied in her nonchalant way, and airily held her telephone out of the window so the BBC audience could hear the full extent of the noise and confirm the sounds of shelling. It rather confirmed Kathy as the intrepid character she was, while we listened sitting in our homes and offices.

[1] Kathy says in her reported acceptance speech when receiving her honorary degree from the University of Zagreb, she recalls that she was awarded the post of Honorary Private, though more widely she is thought to have been made an Honorary Colonel.

Next to the Pile harbour area, our group turned to walk about a very short street or square, with some four dwellings. Remarkably the City Council decided to honour Kathy in recognition of her many contributions and her work with the IUC, by installing a memorial plaque here, recognizing her as an honorary citizen of the city of Dubrovnik, for "her outstanding friendship and courageous support during the 1991–95 aggression". Later that street would also be named after her.

At the unveiling ceremony in February 2012, then Mayor of Dubrovnik, Mr. Andro Vlahušić, and the Chair of the City Council, Ms. Olga Muratti, expressed their deep thanks to Professor Wilkes for her dedicated help to Dubrovnik during the war period in 1991 when she refused to leave the city while electricity was cut off and there were food shortages. She also used her many connections through interviews and reports, and brief forays abroad, to inform the world about war damages, and to seek support and medical resources. She even raised funds to pay for mine removal equipment. In 2001, she would be recognised further, with the award of an honorary degree by Zagreb University.

Her decision to stay in Dubrovnik during the brutal attack on the city, "to be with her brave friends", the citizens of Dubrovnik, to share their ordeals first hand, has left a deep impression. So, it really was a moving moment for all our memorial conference participants to gather under the plaque for various memorial photographs of our own.

Next, our band moved through the busy Stradun or main drag of this fine medieval city, which was hit during the siege, taking in the bustle and hustle of the many wide-eyed tourists, until we reached the City Hall, where in the main assembly room there are just four framed portraits of honoured dignitaries. These included Stjepan Stipe Mesić, Croatian lawyer and politician, who served as President of Croatia (2000–10), and before that prime minister of Croatia (1990).

Alongside is a portrait of Lord (Christopher) Patten, currently Chancellor of the University of Oxford, but formerly the European Commissioner for External Relations and so much involved in the formation and establishment of the new republics that emerged from the former Yugoslavia, including an independent Croatia.

Next, there is the founder of the IUC itself, Professor Ivan Supek, to whom it would also fall to supervise the rebuilding and restoration of the Centre after it was bombed and the library burnt in the 1991–2 siege and war. Harrowing images of the devastation can be seen in record albums when visiting the Centre.

The IUC had been founded 20 years earlier following an international gathering of university leaders held in Montreal in 1970, when the Rector of Zagreb University appealed to his colleagues to help him to create a new kind of Peace University that would be "free of government control" and which could serve as a meeting place for East European academics to meet in dialogue with colleagues from West Europe. He managed to gain the support of some 250 university leaders and

was given the old Teachers' College in the city which he turned into the IUC, along the way persuading Norwegian sociologist/mathematician, Johan Vincent Galtung, who was promoting a discipline of Peace and Conflict studies, to serve as a founding director.

By the time Kathy joined the centre in the 1980s, alongside her Oxford colleague, Bill Newton-Smith, it was already renowned as a cross-over point—allowing dissident intellectuals from the former East Central Europe region to meet with sympathetic Western liberal, anti-authoritarian thinkers, sharing ideas, building moral and generating solidarity, of the deep, lasting, personal kind.

It is acknowledged that embryonic plans for a new free, model, international university for the region were first floated at these IUC Dubrovnik meetings—which later emerged in 1991 in Prague as the Central European University, with an outpost in Warsaw. Its later homes in Budapest, and Vienna, now its current main home, were also explored at the time, as well as Bratislava. As its founding CEO Secretary-General, I was involved in all these discussions and regard the IUC with due reverence.

Finally, there is the portrait of Professor Kathleen Wilkes, who appears to have been on a long journey from her relatively traditional English roots to her position as honorary citizen of Dubrovnik. Yet, perhaps, it was not quite as surprising as it might seem. She was born into the Anglican faith, her father was a vicar in the Church of England, yet she gave up religion. Both sides of her family appear to have been connected: her father taught at the renowned English public school, Eton, and her mother also had connections to the school. Eton is characterised for its apparent elitism, but it seems to have left Kathy with the will and desire to stand up for the underdog and fight for access for the disadvantaged. She would also give up family sympathies for the Conservative Party, and emerge with strong liberal, and, at times, even socialist, sympathies—but never, of course, for Communism as to be found in the Soviet bloc.

It strikes me that there are at least three pathways by which readers of this volume can find connections to Kathy—and, indeed, share in feelings of gratitude for all that she achieved in general, and for the IUC in particular. First, of course, Kathy as the Philosopher of Science. She is probably best remembered for her many contributions to the Philosophy of Mind, especially on the so-called mind-body problem. Her many articles and chapters, more than 50, in professional journals established her reputation as a leading exponent of a nuanced and realistic view of physicalism, that there is just one real or physical world. Her two main works were *Physicalism* (1978) and *Real People* (1988).

She was influenced by many leading thinkers at Oxford where she studied classics (or "Greats"), at St Hugh's College, before going onto to spend three years at Princeton, where she studied with Thomas Nagel, and Richard Rorty among other highly distinguished figures, then to

King's College, Cambridge, and finally settling at St Hilda's College, Oxford, where she was a Fellow in Philosophy for 30 years, and interacting actively with many distinguished fellow philosophers, but also in many other disciplines. Indeed, her willingness to engage directly with scientists and operate freely in an inter-disciplinary frame, particularly with medics, physiologists and psychologists, marked her out. She taught philosophy of science, especially brain and behavioural sciences, but also ancient philosophy, philosophy of mind, and philosophy of religion.

A sense of her characteristic exuberance and down-to-earth approach is, in fact, so well described by Professor Denis Noble in his paper, on Teleology, included in this volume. Denis Noble discusses the lively exchanges that took place in seminars on *behaviourism* he shared with Kathy and others in the 1980s, which would ultimately find exposure in a book entitled *Goals, No Goals and Own Goals.* Kathy is shown as a philosopher, looking for reasonable explanations, appreciating complexities over simpler models often favoured by biologists and behaviourists such as Watson, and who veered towards practicalities and away from "grand theory" approaches.

At St Hilda's, where she lived almost continuously in college in a variety of rooms, she became what we would term a fixture, as a well-known and much loved, if somewhat formidable, figure around the college, dining in hall, lecturing at the University, spending days in the Bodleian Library. Above all, she developed into a hugely conscientious tutor—and this also I know first-hand from many close contemporaries—which earned her much respect and affection, not just from her many students but also from her academic colleagues. They report on stimulating and rigorous tutorials, but also on her patience and her caring attitudes. She always seemed ready to give and share: I can still recall, particularly, one evening with her, encouraged by some good red wine, where she patiently and excitedly explained to me the significance of the Stoics, whom she much admired.

As the "good citizen", she was also a feisty College and University committees member—for example holding her ground on pressing her minority position for her College, the last of the Oxford colleges originally founded for women students—to embrace the brave new world and "go mixed" (as it is now, of course), pressing the Philosophy subfaculty to support the initiative to support Czech dissident academics, discussed below, and, as we noted, holding out at the IUC despite the clear personal danger of falling ordinance round and about.

Second, so pertinent for this volume, we can recognise Kathy as a stout defender of academic freedoms, who absolutely put her principles into practice. Indeed, she was rather proud to be descended from the great English eighteenth century liberal defender of free speech and freedom of the press, John Wilkes.

Kathy, encouraged by her good friend, Steven Lukes, the Oxford sociologist, appears to have played a critical role, as the secretary of the Oxford Philosophy Sub faculty, in persuading it to respond seriously and positively to what appeared as a much corrected, typed askew, somewhat crumpled, letter, from the Czech philosopher, Julius Tomin, appealing to four apparently leading world universities to uphold their responsibilities to protect academic freedom by coming to Prague to continue teaching philosophy to those who still wanted to study and learn, even though the Communist regime had expelled many academics and students preventing them from pursuing studies in such "banned" topics as Kant's ideas, structuralism, phenomenology.

As it turned out, Oxford was the only one of the four approached universities to respond—the others I believe were the Free University in Berlin, Harvard, and the Sorbonne. Kathy herself was one of the first of the so-called Oxford *Velvet Philosophers* to visit Prague to deliver "underground" lectures which were hosted in basements, flats, and other private spaces. Kathy's first visit resulted in four long seminars which the crowd was eager to hear and study, but after two more visits, she was met at the airport by the FSB, the Czech secret police based in Bartolomějská Street, and she was turned back even before starting her talk, and other Oxford philosophers were also expelled.

These expulsions, though, received world-wide coverage, and led directly to a group of largely Oxford philosophers creating the Jan Hus Educational Foundation, which was committed to send regular lecturers and supporters to give talks and lectures at private seminars, provide material support and smuggle *samizdat (clandestine)* books sometimes disguised as novels, and bringing back dissident writings. Among this group can be numbered Alan Montefiore, Ralph Walker, Roger Scruton with many other supporters. Our Jan Hus group would often be working with leading members of the Human Rights, *Charter 77* group, that was founded in 1977 by Vaclav Havel and others in the wake of the famous Helsinki Accords on human rights—including academic freedom—in 1975.

Kathy's continuing support for the JHEF and for Czech dissidents would later earn her a Commemorative Medal in 1998 from Vaclav Havel as President of the Czech Republic. She would make several visits to Prague, made many friends. She had an ungainly walk and was often in pain, having had her back damaged in a riding accident in her teens when her horse tripped over an unseen wire, and yet she would walk the streets until her limbs and back ached in her efforts to lead any following secret police on wild goose chases. In 1981, she volunteered to drive Julius Tomin and his family out of Prague across the border, with risk, ultimately, to safety—and, indeed, freedom—in Oxford, where she went on to help ensure their housing and schooling for the two sons, and the family settling in the UK with Zdena Tominova, former Charter 77 spokesperson going on to work for the BBC's World Service Czech bureau.

All this rather confirmed her status as *persona non-grata* and denied visa access to Czechoslovakia. But this merely seems to have led Kathy to extend her many contacts behind the Iron Curtain and across many countries in Soviet-controlled Eastern Europe beyond Dubrovnik and Prague. Kathy was soon involved in meeting a wide range of beleaguered contacts, from Poland and Bulgaria, mainly, but also Romania, Albania and then the former Yugoslavia, often taking along colleagues and students but also trying to get them invitations to the West.

Indeed, one could say that this became her politics. Not the formal politics of parties and votes, manifestoes and parliaments, but a kind of politics of emancipation, of supporting people, intellectuals, academics, trying to live out their lives openly, trying to think freely, and to realise their own potential.

A third frame is Kathy as friend to so many. For Kathy, everyone mattered. Her students of course, as discussed above, her colleagues, and her fellow philosophers. She loved people, she was warm hearted, and she was always ready to put herself out. She was generous with her time and always ready to share her ideas. Whenever I met her in the early 1980s, I found she had made a point of reading my most recent article so we could discuss it. When I pointed out politely that this was really not expected or required, she would look at me quizzically, and, rather kindly, said she would want to read them anyway, it was important.

Val McDermid, the well-known crime novelist, who was taught by Kathy, also remembers her as a woman of "great generosity and sheer brilliance". She recalls often discussing philosophy with Kathy long into the night. "Kathy taught me how to think, and she also taught me how to drink!" she added. That could be true for many of us. Another close friend who met and go to know Kathy in from Prague, Jana Frankova, recalls the warm personality of Kathy who "as a friend was open, sincere and helpful' and who became like a member of her family", helping to look after her kids and take them shopping while smuggling in books and money. The fact that Jana too was a St Hilda's language graduate no doubt helped. She recalls a personal friendship that "made the lives of so many people in our country at least somewhat easier". It mattered that those from the UK and other western European countries came and cared.

Val went on to base a key character in her novel, *The Skeleton Road*', on Kathy set during the siege of Dubrovnik. "She considered it her duty to help and support the people of the city and to inform the world what was happening there," she says.

She would never make much of her work supporting dissidents, which she considered, disconcertingly perhaps, as "normal". She could do it, so she did, and so would anyone else, with the time and the resources, perhaps the understanding, who thought deeply enough about it. Of course, this is a fallacy. But it helps reveal much of the nature of Kathy—the deeply liberal-democratic global citizen.

She was meeting people who could not do what she took for granted—to be able to read, think and discuss. Surely, such individual acts would pose no clear political threat. Yet, of course, it challenged the underlying ethos of the Soviet regimes. The Communist party elites were fearful about where this all might lead, no doubt influenced by how similar "freedoms" generated the momentum that led to the 1968 Prague Spring, which ended in the infamous USSR tank invasion.

All these three frames pointed the way towards Kathy emerging, somewhat inevitably, as what might be termed a global citizen of city of Dubrovnik. She joined the board of the Inter-University Centre in the early 1980s, and became Chair of its Executive Committee in 1988, until 1996, though staying in touch right to the end of her days.

She is remembered as an "outstanding", "active", and "innovative" Chair, always at hand to give advice, and utilize her wide network of contacts with prominent scholars and sources of financing to support the IUC.

She would play a key role in many of its academic programmes, but especially in the Philosophy of Science seminar series, to which she invited and brought in many distinguished colleagues from all over the world, including some involved in this volume, and which has been held every year since its initiation. This volume celebrates the return of what the IUC, understandably, regards as one of its most prominent course series after the Covid-19 interruption.

Yet, all these activities had taken their toll on her health and personal well-being. In 2003, I can remember a delightful sunny evening with a few other friends and former students, sharing some red wine on the balcony of her rooms in St Hilda's, overlooking a batch of then current students practicing sports in the evening sunshine. As we reviewed past exploits, laughed over memories, recalled common friends, delighted in the peacefulness and idyll of the scene, all seemed right with the world. To our great sadness, though, she would die not long after.

Freedom, people, philosophy all mattered to Kathy. Above all it was Philosophy which opens the mind to new ideas and new ways of thinking, a kind of liberation, which in many ways, serves as a fine metaphor for the work of the Velvet Philosophers generally, and Kathy, in particular, crossing borders in spite of restrictions rules imposed by Communist regimes, to help unlock these new modes of thinking. The IUC was founded on the principle of international "openness" and Kathy more than fully subscribed this principle.

None, by definition, were or should be contained, within boundaries, or limits. She associated so closely and personally with those brave Prague dissident philosophers and intellectuals she first met in the 1980s, and she carried her commitment to them and their cause, almost with religiosity, to other Soviet territories and ultimately in support for the IUC and Dubrovnik.

This conference, represented in this volume, was the second in what is planned as a regular series around themes in the philosophy of science. The first was held in Oxford at St Hilda's, as part of the College's 125th anniversary celebrations. It was inspired by Professor Anita Avramides, who worked with Kathy for three decades at St Hilda's and covered for her when Kathy opted to stay on in Dubrovnik during the siege. She was supported by Professor Riccardo Viale, Director of the Herbert Simon Society based in Torino, Italy, who had studied at Balliol College, Oxford, with Professor Bill Newton-Smith. He also interacted directly with Kathy, who had first invited him to an IUC conference in 1983 even though he was, at that time, a psychologist. It was an experience that convinced him to switch his interests to philosophy, full time. Finally, myself, as a close friend of Kathy from my Oxford student days, but more as a colleague within the Jan Hus's various Prague initiatives, and as a journalist reporting the work with dissidents. Nada Bruer, the secretary of the IUC, was quick and enthusiastic to join the initiative once she was approached, and the conferences series was set up with three pillars linked to Oxford, Dubrovnik, and Torino.

The Dubrovnik conference allowed us all to share new thinking on a range of the themes within the Philosophy of Science. It also allowed us to recall and share memories of Kathy Wilkes. This volume is dedicated to her in the name and to new thinking.

Croatian Journal of Philosophy
Vol. XXII, No. 66, 2022
https://doi.org/10.52685/cjp.22.66.3
Received: August 15, 2022
Accepted: September 15, 2022

Kathy Wilkes, Teleology, and the Explanation of Behaviour

DENIS NOBLE*
University of Oxford, Oxford, UK

Kathy Wilkes contributed to two books on Goal-directed Behaviour and Modelling the Mind based on interdisciplinary graduate classes at Oxford during the 1980s. In this article, I assess her contributions to those discussions. She championed the school of philosophers who prefer problem dissolution to problem-solution. She also addressed the problem of realism in psychology. But the contribution that has turned out to be most relevant to subsequent work was her idea that in modelling the mind, we might need to "use as structural elements synthetic cells, or things that behaved very like neurones." I show how this idea has been developed in my own recent work with zoologist and neuroscientist, Raymond Noble, to become a possible physiological basis for the ability of organisms to choose between alternative actions, and so become active agents. I consider that this insight became her seminal contribution in this field.

Keywords: *Teleology; goal-directed behaviour; modelling the mind; agency*.

1. *Introduction*

It was a great privilege for me to give the opening lecture at the Dubrovnik Inter-University Centre symposium honouring Kathy Wilkes.[1]

* I thank Anthony Kenny, Alan Montefiore, Andrew Packard and Raymond Noble for many discussions that have contributed to my thinking about this subject. Andrew Packard was particularly helpful in drawing my attention to multiple aspects of the work of JZ Young.

[1] This Conference was held at the Inter-University Centre, Dubrovnik, 29th April 2022.

This article is closely based on that lecture. My main credentials for doing so arise from seminars held in Balliol College Oxford during the 1980s on the explanation of animal and human behaviour. They arose from a long-standing collaboration between Alan Montefiore, a philosopher, and me, a biological and medical scientist.

We both edited the book *Goals, No Goals and Own Goals* (Montefiore and Noble 1989) that resulted from the seminars with Kathy Wilkes, and David McFarland, an ethologist, as co-organisers. Alan's description[2] of the way the debate developed is correct when he says that there was mostly an axis between Alan and me on the one hand and one between Kathy and David on the other. This outcome is itself significant. The divide was not really one between scientists and philosophers, and it shows also that scientists themselves are not neutral with respect to philosophical concepts concerning animal and human behaviour. There were two other contributors: Shawn Lockery, now a Professor of Neuroscience in the USA, and Dan Dennett, who contributed an article but did not take part in the seminars.

A further professional link with Kathy arose from the book she edited with Bill Newton-Smith, *Modelling the Mind* (Said, Newton-Smith et al. 1990). We both contributed chapters to that book. Kathy herself wrote the chapter (Wilkes 1990) that gave the book its title, while I followed some of the arguments in the *Goals* book, with a chapter on Biological Explanation and Intentional Behaviour (Noble 1990).

2. *The philosophical and scientific background*

My own interest and involvement in these seminars arose from a published interaction in 1967 with the Canadian philosopher Charles Taylor, following his book *The Explanation of Behaviour* (Taylor 1964), based on his doctoral thesis at the University of Oxford. I was introduced to the book by Anthony Kenny, who was working on related problems (Kenny 1969), and with whom I have interacted ever since on issues to do with mind, will and action. Arising out of our discussions he encouraged me to write a critique of Taylor's book, which was published in *Analysis* (Noble 1967), where I argued that Taylor's defence of teleological explanation was incorrect since it seemed to require that a difference in state at one (higher) level should not necessarily have a correlate at another (lower) level. On this view, there would be a gap in the mapping. As a physiologist I found the idea of such a gap difficult to accept.

Taylor did however reply with a very interesting argument (Taylor 1967). This was that, while there could not be a physical gap it might nevertheless be the case that, after studying a whole series of correlations between, say, behaviour and neural states, only the higher level of behaviour might show a pattern that could count as an explanation.

[2] Personal communication.

Specifically, if the behaviour states are B1, B2, B3, … and the neural event states E1, E2, E3…. the E states might be disordered with respect to explaining the behaviour whereas the B states might offer a ready explanation. I found this a very interesting reply and countered that the consequence was that the issue of the validity of teleological explanations became a conceptual issue, not an empirical one (Noble 1967). I believe that was an important clarification, and that it is still valid. The clarification will reappear later in this article. But I also think the debates have moved on very significantly since 1967.

The seminars in Balliol in which Kathy was such a major contributor formed an important stage in that development. During those seminars I was still developing the ideas on goal-directed behaviour that eventually became expressed in my more recent books *The Music of Life* (Noble 2006) and *Dance to the Tune of Life* (Noble 2016) and even more recent articles (Noble 2017, Noble and Noble 2017, Noble and Noble 2018, Noble, Tasaki et al. 2019, Noble and Noble 2021). Those publications describe the ways in which teleological behaviour naturally occurs and develops during the evolutionary process. They also show how such behaviour itself contributes to evolution and so gives evolution itself a kind of directionality. Most recently, these include a paper on purpose in physiology appearing in the *Biological Journal of the Linnean Society* (Noble and Noble 2022). I will return to what led to those books and articles at the end of this article, by showing how one of Kathy's contributions formed a key element in those developments.

However, I was far from ready during the Balliol seminars in the 1980s to give expression to those ideas at that time. It is only in retrospect that I can see the roots of my development. That is unfortunate from one point of view. If I had been able to express the ideas and marshal the biological experimental evidence more forcefully in the 1980s perhaps the debates in which Kathy was involved would have taken a different turn. But the flip side of this coin is that I remain deeply grateful to Kathy herself, and to the other participants for a sustained and deeply stimulating series of seminars that did much to clarify my own thinking. I would have loved to try the more recent ideas out on Kathy, particularly because, as I will show, I believe they answer one of the key questions she contributed to the debates of the *Goals* book.

3. *Reactions to the book*

Soon after publication of the *Goals* book in 1989, I sent a copy to the distinguished zoologist and expert on the intelligent behaviour of the cephalopods, JZ Young. I had been taught medical sciences in UCL where he was the professor of Anatomy and a world-renowned expert on the learning and behavioural repertoires of the octopus. I suspect I learnt more philosophy from him than anatomy! So, it seemed a good idea to get his reactions. He wrote to me afterwards to say that he had enjoyed reading it, several times in fact. But he wasn't exactly com-

plimentary as far as my own contributions were concerned (still the critical professor of his former student!) and he didn't seem to go much for Alan's contributions either. So much for Alan's and my side of the debate! But JZ Young was *much* more complimentary about Kathy's chapters which he thought were clear and, in his view, largely correct.

Why was JZ Young sympathetic to our debate at all, and to Kathy's contributions in particular, even though critical of some of what Alan and I wrote? To understand that we need to recall that JZ Young was the discoverer of the giant nerve axon in the squid (Young 1936, Young 1938, Keynes 2005) that enables it to trigger a form of jet propulsion (Packard 1969), in turn enabling it to successfully flee predators. This was the giant nerve on which Alan Hodgkin and Andrew Huxley worked to obtain the experimental data on which they constructed their famous mathematical model (Hodgkin and Huxley 1952) of the nerve impulse and its dependence on sodium and potassium ion channels in the axon membrane. It was an important prediction of their model that large nerve axons would conduct faster than small ones, as they were known to do (Pumphrey and Young 1938), though it should be added that this was not the reason for their choice of nerve to work on. The squid axon was simply large enough for them to insert their recording and controlling electrodes. When Hodgkin and Huxley were awarded the Nobel Prize for this work in 1963 Young was known to have commented that this was a bit like awarding a prize to the typewriter rather than to the book author. I don't think he meant to denigrate Hodgkin and Huxley's achievement. Rather he was pointing out that the reason for the existence of the giant axon, its purpose, was the evolutionary imperative to generate a rapid response to predators. Furthermore, the giant axon was not an evolutionary development found in all cephalopods. It is not found in octopods. The efficiency of the jet propulsion mechanism depends therefore more on the functional anatomy of the whole system ensuring simultaneous contraction, not just the speed of nerve conduction. He saw that this was the emergence, during evolution, of a goal directed mechanism. Every aspect of the anatomy and physiology of the cephalopods was fine-tuned in ways that endowed the organisms with a rapid escape mechanism.

He therefore regarded the mathematical analysis of the mechanism of the nerve impulse to be too low a level to explain the goal-directedness of the behaviour, with which I am sure Hodgkin and Huxley would have agreed. So, he was certainly sympathetic to the general purpose of the *Goals* book. Low-level explanations don't work, and for precisely the reason that emerged from my interaction with Charles Taylor. Incidentally, there is a very useful "Celebration of JZ Young" by Andrew Packard and Fabio DeSio published in *Physiology News* in 2010 (Packard and DeSio 2010). I see JZ Young as the embodiment of the tension between purposive and reductive accounts of biology, a view that is reinforced by this quotation from one of his collaborators, Brian Boycott:

> there is, in most of JZ's scientific design and output, a tension between his desire to investigate integrative functions of organs and systems as a whole and the practical constraint that to do this requires the reduction of a system to an experimentally manageable and interpretable entity. (cited by Packard and DeSio 2010)

So, why did JZ Young think more of what Kathy wrote than what Alan and I wrote? I suspect that he was nevertheless suspicious of teleological ways of speaking about animal behaviour. Most biological scientists were sceptical of that approach in the mid-twentieth century: "Teleology is like a mistress to a biologist: he cannot live without her but he's unwilling to be seen with her in public."[3] Some biologists even invented the word teleonomy (Pittendrigh 1958) to refer to the biological processes involved without committing to whether or not an organism is an active agent.

I now think that there was no need to invent a separate word. Organisms are definitively purposeful agents (Noble and Noble 2022). But this is not the place to justify that point. Here it suffices to say that it is a tribute to Kathy's work that such a noted expert on animal behaviour as JZ Young thought highly of it. So, what were the main points of her contributions to the *Goals* book?

4. *Kathy's contributions*

She wrote two chapters in the book, and she explains her philosophical position most clearly at the end of the second (210). She wrote:

> Our discussions of these issues over several years have left me more confused at the end than I was at the beginning.

(Surprise, surprise!) …. And then continues

> I have suggested that many of the problems might be pseudo-problems—to be dissolved rather than solved; certainly I align myself with the 'theft over honest toil' school of philosophers who prefer problem dissolution to problem-solution.

Nevertheless, she identified

> one question [that] has emerged as indissoluble, crucial and critical: what counts as 'realism' in psychology? This needs serious thought, which would and should enrich and deepen the ongoing examination of realism in the physical sciences.

On this, she was surely right. There is a veritable flood of books now on *What is Real* (Becker 2018), *The Matter with Things* (McGilchrist 2021) and similar titles, to which I would add Hilary Lawson's groundbreaking analysis of "reality" in his book *Closure: A Story of Everything* (Lawson 2001). As I will show at the end of the paper, there are good reasons for this explosion: there is a groundswell of opinion in opposition to the confidently-expressed materialist (realist) certainties of the mid-20th century.

[3] Attributed to J. B. S. Haldane.

Kathy herself was more concerned with what the common "man in the street" might want as explanations. She wrote:

> not all explanations are causal explanations…if one job of explanation is to remove puzzlement, then evidently people can be puzzled by well-nigh anything.

Here she is talking very much in the tradition of philosophy paying attention to the language of the man in the street. I found it helpful that she kept bringing us back to the pragmatic uses of philosophy. This aspect of her work was, I suspect, at one with her engagement with the problems of the world, notably here in the immense contributions she made to the cultural life of Dubrovnik, and of course the amazing contribution she made to intellectual life in Prague. There are others at this symposium who know far more than me about that aspect of Kathy. My knowledge is second-hand, largely through two other Oxford philosophers, Bill Newton-Smith and Anthony Kenny, who both lectured to the under-cover seminars held in Prague. In a recent email to me, Kenny writes:

> When the dissident Czech philosophers first made contact with Western Universities, only Oxford made a positive response, and that was due to Kathy who was then secretary to the Philosophy sub-faculty. I think that she, Bill Newton Smith and I were the only people to be arrested for talking to the Tomin group—but it was she who went on lecturing after being arrested. Nancy and I were just taken off to the police station and extradited early the following day (to the surprise of the German frontier police who assumed we were drug smugglers).

Time and again, Kathy was concerned more with pragmatics than with grand theory, of which it seems to me she was highly sceptical. By contrast, Alan and I must have seemed to her to be too strongly concerned with conceptual theory.

In this vein, here is what she thought about whether science could find correlates of intentions:

> [Common sense psychology] needn't bother about whether these intentions are explicit and real, or tacit and hence not really 'there' in any strong sense. In other words, when we ascribe intentions to an agent, we are not usually …committing ourselves to the existence of a physical correlate *to that very intention*.

I suspect that this is why she and David McFarland often joined forces. David, as an ethologist, was very sceptical of whether intentions matter at all! If I understood him correctly, these were feelings we experience but which need not have any influence on how we actually behave.

Kathy herself was not, of course, a Watsonian. She writes:

> Extreme (Watsonian) behaviourism failed because there is so much that it just cannot explain. This is scarcely surprising; it always was a priori implausible that so simple-minded a theory could account for the most complex system we know. But it rejected all 'mental' terms; here I am only examining the possibility that a scientific theory might do without one of them: intentions.

A strong feature of Kathy's contributions to the debate was her continual insistence on clarifying what we mean by an explanation:

> What sort of 'accounting for' [do] we want 'the traditional goal concept' to provide… In this book we find free use of 'causes', 'is responsible for', 'explains' 'continually guides' and more besides. This leads into the rather more specific question of whether explanation via intentions, or goal representations, is a species of causal explanation. And that forces one to ask just what is needed if A is to be 'the cause' of B: 'being a cause', 'serving to explain', and 'being responsible for' are not synonymous expressions. (195)

On this issue, Kathy and I were in agreement. We both thought that, whatever intentions might be, they could not be the cause of behaviour in the same kind of way in which nerve action potentials cause muscle movement. I think she was on exactly the right lines in insisting that, at the least, different concepts of cause need to be invoked. She wrote:

> Thus, although endorsing Noble's claim (97) that 'within an intentional context a "machine" description of what happens fails to make reference to the most significant facts' I would want to explain why this must be so by linking 'significance' to the precise characterisation of the explanandum—to the puzzlement of the inquirer. I find it increasingly difficult to find any real-life cases where there is genuine, honest-to-goodness' 'rivalry' at all between intensional and non-intensional explanations of what is indeed the one and the same explanandum.

These arguments all form part of her "attempt to underline the differences between common sense, and scientific, explanation" (198).

I now find myself in total agreement with her arguments on this issue, even to the extent that my own recent publications not only elaborate on why intentions cannot be causes in the same way as nerve impulses can be, but also that, even within purely biological levels of organisation, the forms of cause between different levels can be quite different. As an example, causation from the genetic level is mediated by templates (gene sequences) not by specific molecular interactions (Noble, Tasaki et al. 2019).

These direct quotes from her work for the *Goals* book will, I hope, give readers a flavour, at least, of what Kathy contributed to the seminars and the book. Fortunately, the book itself has been republished as an e-book by the publisher (Montefiore and Noble 2021), so interested readers can readily explore further if they wish.

Now I turn to her contributions to *Modelling the Mind*. I am not surprised that it became the overall title for that book. For, by contrast with the *Goals* book, where she says herself that she was left more confused, her chapter in the 1990 book represents Kathy in full flow as the insightful philosopher she clearly was.

She begins by clearly stating that we should never talk about *the* model. Even in physics, we need multiple models, even incompatible models, for models, like metaphors, illuminate different aspects of reality, and they can be useful even when incompatible. As Lakoff and Johnson famously said in their 1980 book, *Metaphors We Live By* (La-

koff and Johnson 1980), metaphors can have good and bad ranges of applicability. What works for the micro-level in physics, i.e. quantum mechanics, does not cover what the theory of General Relativity covers and vice versa. She writes:

> The danger, as far as psychology is concerned, comes when we switch from indefinite to definite article. (63)

Yet, particular models do become dominant:

> Hume's metaphor of the mind as an inner theatre was never more than that, a metaphor (as he was the first to insist), even though it became deeply compelling to treat it *as if* the mind were indeed really like that. (64)

So, if we "cannot think of minds as inner theatres, inspected by an unblinking inner eye, any longer" just what do we think the mind might be, or what is it to be mental?

There is then a careful analysis of the computer model of mental processes. She points to the danger that

> there is a real possibility that psychological explanations might 'bottom out' in hardware structure and function long before we have learned anything from the computer metaphor; in fact, that the really interesting work may come rather from one or other of the neurosciences than from simulation exercises. (73)

It is at this point that I encountered a fascinating speculation:

> It may be that if we were to construct a computer with capacities close to those of the human brain, we would have to use as structural elements synthetic cells, or things that behaved very like neurones—with, say, action potentials, graded potentials, 'synaptic' modifiability, 'dendritic' growth, etc. (73–4)

This paragraph is tantalisingly close to where my own thinking has gone recently. Specifically, I have speculated that, in order to access the kinds of molecular stochasticity in real brains, we might have to make "computers" using water rather than silicon. The argument is simply that novelty, creativity, in organisms may depend on precisely what kind of stochasticity is harnessed by living organisms.

My overall conclusion from re-reading Kathy's work after about 30 years have passed, is that her contribution to *Modelling the Mind* is the better example of her thinking. She was in full control of what she was writing, instead of being "more confused at the end than I was at the beginning".

I suspect that one of the reasons for that conclusion on her part is a fault of my own as the biologist in the debates. Perhaps something was missing from what I, as the physiologist, should have contributed.

5. *What was missing in the 1980s?*

I will therefore explain what I believe was missing on my part, at least, during those debates in the 1980s. So, this article now becomes a kind of *mea culpa*. The problem is actually very easy to explain. Like most bio-

logical scientists I was still under the sway of a seminal book, written in 1944 by the great quantum mechanics pioneer, Erwin Schrödinger, called *What is Life?* (Schrödinger 1944). I call it a seminal book because it led to the central Dogma of Molecular Biology in the work of Watson and Crick when they unraveled the double helical structure of DNA. Both acknowledge Schrödinger because he made two predictions in his book that were, apparently, to find their confirmations in the work of Watson and Crick. The first was that the genetic material, when it was discovered, would be found to be what he called an aperiodic crystal. If you think of a linear polymer as a kind of crystal—a bit of stretch, I agree—the description aperiodic is a very good one. It is precisely that characteristic that enables the molecular thread to encode so much information that enables a vast range of different proteins to be constructed by the living cell.

So far so good. But the second prediction of Schrödinger to be taken up by the molecular biologists simply cannot be true. He argued that, if one sees the genetic material as an information dense sequence, how is it read to enable the characteristics of an organism to be transmitted from one generation to another? A one-dimensional sequence cannot simply map a three-dimensional structure. It is not a miniature organism in the way in which some nineteenth century microscopists imagined when they looked at sperm and egg cells. Could that three-dimensional template come from somewhere else, perhaps in the three dimensional structure of the cell itself? Whichever way that is done, Schrödinger reasoned that the sequence must be read in a determinate manner if it was faithfully to transmit information. Stochasticity in a communication line is intolerable. From this he concluded that there must be an absolutely fundamental difference between physics and biology.

Physics can be characterised as order from disorder. At the micro level, there is the essential stochasticity of quantum mechanics. Even if, one day, an alternative view of "reality" is produced, as people like Albert Einstein and David Bohm believed, we can't escape the fact that the equations of quantum mechanics are precisely predictive as probabilistic descriptions. Any underlying determinism would have to reproduce this. That is not difficult to imagine since we already have an example of stochasticity at the molecular level that was discovered well before quantum mechanics. In 1827 Robert Brown observed that fine particles derived from pollen grains showed stochastic movement in water observed under the microscope. We call it Brownian motion and it was shown by Einstein (Einstein 1905) to arise from the random bombardment of the particles by the random motion of water molecules: the first demonstration of the existence of individual molecules with separate motions.

Yet, the equations of thermodynamics, which describe large numbers of particles to generate the gas laws, are determinate. The an-

swer to this apparent paradox is that, if motion at the particle level is genuinely random, then large numbers of particles will cancel their individual movements out to produce a constant pressure when hitting an object, like the wall of a pressure vessel. Order at large scales therefore arises from disorder at lower scales. In a living cell, the high-level properties of volume, pressure, temperature, acidity, and many other global parameters will display constant or smooth transitions.

But this interpretation is inconsistent with a Schrödinger view of biology in which the genetic material at the molecular level is supposed to be read in a determinate manner, rather as an X-ray beam can generate an accurate and determinate "picture" of a crystal by the diffraction of the rays by the regular structure of the crystal. Biology, he reasoned, was therefore the generation of order at large scale from order at the micro scale.

Schrödinger wrote:

> We seem to arrive at the ridiculous conclusion that the clue to understanding of life is that it is based on a pure mechanism, a 'clock-work'…The conclusion is not ridiculous and is, in my opinion, not entirely wrong, but it has to be taken "with a very big grain of salt" (1944: 101).

He then explains the "big grain of salt" by showing that even clockwork is, "after all statistical" (103). My reading of these last pages of Schrödinger's book is that he realises that something is not quite right but is struggling to identify what it might be. This confusion has muddied the waters for 80 years now.

We would now say that the molecules involved (DNA) *are* subject to statistical variation (copying errors, chemical and radiation damage, etc.), which are then corrected by the protein machinery that enables DNA to be a highly reproducible molecule. This is a three-stage process that reduces the error rate from 1 in 104 to around 1 in 1010, which is an astonishing degree of accuracy. The order at the molecular scale is therefore actually imposed *by the system as a whole.* This requires energy of course, which Schrödinger called negative entropy. Perhaps therefore this is what Schrödinger was struggling towards, but we can only see this more clearly in retrospect. He could not have known how much the genetic molecular material experiences stochasticity and is constrained to be highly reproducible by the organism itself.

So Schrödinger's idea that led to the Central Dogma can't be correct. It also led to the incorrect "read only" view of DNA.

Now, why is this important to the debates on teleology? The answer is that the Central Dogma should no longer be used to justify a closed determinate nature to biological processes. Just like everything else that depends on the motion of molecules, there is massive stochasticity at the lowest levels. Only at higher levels can there be the order that a genuine explanation of behaviour requires. Furthermore, it is precisely through the constraints that the higher order imposes on the lower level stochasticity that we can develop a multi-level theory that privileges the

higher level. That is the purpose of two of my most recent articles (Noble 2022a,b) and of my book, *Dance to the Tune of Life* (Noble 2016). Those constraints ensure that there is an asymmetry between the causal force of explanations at higher and lower levels. The higher level is genuinely causative because it is only from that level that one can understand the constraints and how they arise. This is the sense in which I think that Charles Taylor's conceptualist view of teleology is correct, and how I think it can now be given a firm biological science basis.

Furthermore, it is possible to show that this *necessarily* excludes the one-way reductionist causal explanation of organism behaviour. The complete argument is technical, but the overall conclusions are straightforward:

1. When we examine the mathematics of multi-level causation, which is encapsulated in the principle of biological relativity (Noble 2012), it is impossible to dispense with the influences of higher levels on lower-level behaviour. That is a mathematical necessity in any living system in which the molecular level is controlled by higher levels (Noble 2022, Noble 2022).
2. Organisms use lower-level stochasticity to generate their characteristic innovative activity in finding solutions to the challenges of survival. Our immune systems are doing that all the time, and they do so by changing the organism's DNA sequences in a highly targeted way (Odegard and Schatz 2006). That kind of selective targeting was supposed to be forbidden by the Central Dogma. It is not.
3. Similar harnessing of stochasticity occurs in the functioning of the nervous system, so that it becomes possible to explain the physiological processes that might underly innovative behavior (Noble and Noble 2020). It is at one and the same time, both stochastic (we can't necessarily predict a Beethoven or an Einstein), yet understandable in retrospect (we can judge the reasons and values that must have guided what was done).

I therefore think that one aspect of the debate is now closed. Higher level explanations *must* have validity because we cannot dispense with the influences of higher levels on lower-level behaviour. That is a mathematical necessity in any living system in which the molecular level is controlled by higher levels.

I want to conclude by noting that the issues on which Kathy contributed so much 30 years ago are still very much live issues today. If I have succeeded in moving the debate on somewhat I owe a lot to her insights and great contributions. Her insight that we may need to use "things that behave very like neurones" now seems prophetic.

6. *Coda*

Nearly 20 years ago, in August 2003, I was contacted by Alan Montefiore in London to ask whether I could possibly go to the John Radcliffe

hospital in Oxford to visit Kathy Wilkes, who was unwell. I did so. Kathy was indeed unwell. I was trained as a medical student, though I never treated patients, but I was saddened to see all the signs of a hopeless clinical situation. Kathy, though, immediately recognised me and we briefly discussed her work. Her mind was clearly focussed on Croatia and what happened in Dubrovnik. Sharp as a knife, she reacted immediately to my mistake in referring to Yugoslavia (which is what your country was when I first visited it in 1965). I immediately tried to correct what I said, but she was very firm and insistent: what I believe may have been her last words were "I am a fighter, I never give up."

She was!

References

Becker, A. 2018. *What Is Real?: The Unfinished Quest for the Meaning of Quantum Physics*. New York: Basic Civitas Books.

Einstein, A. 1905. "Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen." *Ann der Phys* 17: 549–560.

Hodgkin, A. L. and A. F. Huxley 1952. "A quantitative description of membrane current and its application to conduction and excitation in nerve." *Journal of Physiology* 117: 500–544.

Kenny, A. J. P. 1969. *The Five Ways*. London: Routledge & Kegan Paul.

Keynes, R. D. 2005. "J.Z. and the discovery of squid giant nerve fibre." *Journal of Experimental Biology* 208: 179–180.

Lakoff, G. and M. Johnson 1980. *Metaphors We Live By*. Chicago: University of Chicago Press.

Lawson, H. 2001. *Closure: A Story of Everything*. London: Routledge.

McGilchrist, I. 2021. *The Matter with Things*. London: Perspectiva Press.

Montefiore, A. C. R. G. and D. Noble (eds.) 1989. *Goals, No Goals and Own Goals*. London: Unwin-Hyman.

Montefiore, A. C. R. G. and D. Noble (eds.) 2021. *Goals, No Goals and Own Goals*. London: Routledge.

Noble, D. 1967a. "Charles Taylor on teleological explanation." *Analysis* 27: 96–103.

Noble, D. 1967b. "The conceptualist view of teleology." *Analysis* 28: 62–63.

Noble, D. 1990. "Biological Explanation and Intentional Behaviour." In K. A. M. Said, W. H. Newton-Smith, R. Viale and K. Wilkes (eds.). *Modelling the Mind*. Oxford: Oxford University Press, 97–112.

Noble, D. 2006. *The Music of Life*. Oxford: Oxford University Press.

Noble, D. 2012. "A Theory of Biological Relativity: no privileged level of causation." *Interface Focus* 2: 55–64.

Noble, D. 2016. *Dance to the Tune of Life: Biological Relativity*. Cambridge: Cambridge University Press.

Noble, D. 2017. "Evolution viewed from physics, physiology and medicine." *Interface Focus* https://doi.org/10.1098/rsfs.2016.0159.

Noble, D. 2022. "How the Hodgkin Cycle became the Principle of Biological Relativity." *Journal of Physiology* https://doi.org/10.1113/JP283193

Noble, D. 2022. "Modern Physiology Vindicates Darwin's Dream." *Experimental Physiology* doi: 10.1113/EP090133

Noble, D. and Noble R. 2021. "Rehabilitation of Karl Popper's Ideas on Evolutionary Biology and the Nature of Biological Science." In Z. Parusniková and D. Merritt (eds.). *Karl Popper's Science and Philosophy*. Cham: Springer International Publishing, 193–209.

Noble, R. and Noble, D. 2017. "Was the Watchmaker Blind? Or Was She One-Eyed?" *Biology* 47: 10.3390/biology6040047.

Noble, R. and Noble, D. 2018. "Harnessing stochasticity: How do organisms make choices?" *Chaos* 28: 106309.

Noble, R. and Noble, D. 2020. "Can Reasons and Values Influence Action: How Might Intentional Agency Work Physiologically?" *Journal for the General Philosophy of Science* 52: 277–295.

Noble, R. and Noble, D. 2022. "Physiology restores purpose to evolutionary biology." *Biological Journal of the Linnean Society.*

Noble, R., et al. 2019. "Biological Relativity Requires Circular Causality but Not Symmetry of Causation: So, Where, What and When Are the Boundaries?" *Frontiers in Physiology*: 10.3389/fphys.2019.00827.

Odegard, V. and Schatz, D. 2006. "Targeting of somatic hypermutation." *Nature Reviews in Immunology* 6: 573–583.

Packard, A. 1969. "Jet Propulsion and the Giant Fibre Response of Loligo." *Nature* 221: 857–877.

Packard, A. and DeSio, F. 2010. "Celebration of JZ Young." *Physiology News* https://cephalopod.files.wordpress.com/2017/06/desio-n-packard-pn-78-2010-cover.pdf

Pittendrigh, C. S. 1958. "Adaptation, natural selection, and behavior." In A. Roe and G. G. Simpson (eds.). *Behavior and Evolution.* New Haven: Yale University Press, 390–416.

Pumphrey, R. J. and Young. J. Z. 1938. "The rates of conduction of nerve fibres of various diameters in cephalopods." *Journal of Experimental Biology* 15: 453–466.

Said, K. A. M., et al. 1990. *Modelling the Mind.* Oxford: Oxford University Press.

Schrödinger, E. 1944. *What is Life?* Cambridge: Cambridge University Press.

Taylor, C. 1964. *The explanation of behaviour*. London: Routledge & Kegan Paul.

Taylor, C. 1967. "Teleological explanation – a reply to Denis Noble." *Analysis* 27: 141–143.

Wilkes, K. V. 1990. "Modelling the Mind." In K. A. M. Said, W. H. Newton-Smith, R. Viale and K. Wilkes (eds.). *Modelling the Mind.* Oxford: Oxford University Press, 62–82.

Croatian Journal of Philosophy
Vol. XXII, No. 66, 2022
https://doi.org/10.52685/cjp.22.66.4
Received: June 25, 2022
Accepted: August 16, 2022

Intentions and Their Role in (the Explanation of) Language Change

DUNJA JUTRONIĆ
University of Split, Split, Croatia

The primary aim of this article is to find out what different linguists say about the role of intentions in the study and explanations of language change. I try to investigate if in the explanation of language change, "having an intention" does any explanatory work. If intentions play a role, how do they do it, at which point it is salutary to invoke them, and what do they contribute to the explanation of language change? My main claim is that speakers' intentions have a role to play only on higher linguistic levels, e.i., in speakers' communicative strategies. Since this is a celebration for Kathy Wilkes and her contribution to goal-directed behaviour, in the Concluding remarks I go back to her remarks on language and intentions and see how they fit my discussion in this paper.

Keywords:Language change; speakers' intentions; goals of communication; Kathy Wilkes.

1. *Introduction*

The primary aim of this article is to investigate if in the explanation of language change, "having an intention" does any explanatory work. What I want to find out is if intentions do play a role, how do they do it, at which point it is salutary to invoke them, and what do they contribute to the explanation of language change.

It is crucial for the discussion to make a clear distinction between: (1) doing A intentionally[1] *vs.* (2) having an intention to do A. The follow-

[1] Tomasello says: "So why don't apes point?... they do not understand communicative "intentions" (208: 385); "...only humans have the skills and motivations to engage with others collaboratively, to form with others joint intentions and joint attention in acts of shared intentionality (2008: 387). Tomasello is talking about

ing example shows the difference between the two: He intentionally (1) ran to the station thus causing a heart attack but he did not intentionally (2) cause a heart attack. What is very important in this discussion is that intentionality as intending to do things (no. 1) should not be confused with having an intention to act (no. 2). Namely, intentions in the sense of having a thought to act, or to have a thought about language, to have a thought about reference, etc. are propositions attitudes. Bratman says that one has: to spell out "the relation between intentional action [intentionality no. 1.] and intending to act, i. e. having an intention to act [intentionality no. 2]" (1984: 375). He says: "Intentions are distinctive states of mind" (1984: 376), or as Devitt says: "Intentions, like beliefs and desires are thoughts, propositional attitudes" (2021, on the web). In this article I concentrate on no. 2 intentionality, i. e., on intentions as having a belief/thought about something, here particularly, having a thought about language. I follow the application of this distinction in linguists' writings about language and language change. I ask if "having an intention" (no. 2) plays explanatory work in language change.

I proceed as follows: In section 2 under the subtitle *Causes of language change* I present some old and some more recent opinions on the causes of language change. In section 3 *What kind of "beast" language is?* I set the scene and restrict myself to the discussion of two models of language: language as autonomous system *vs.* language as the "rational agent" system. The question is: Does language change happen internally by itself or do speakers have an important role in language change? In section 4 under the subtitle *Transferring the evolutionary metaphor to the case of language change*, I discuss the adoption and adaptations of the theory of biological evolution as applied to an evolutionary theory of language change and mostly present William Croft's evolutionary theory of language change. The role of intentions stays the central issue. In section 5 under the title *On speakers' intentions* I review what has been said about intentions in language change. Section 6 points to and discusses *Problems with explaining change with speakers' intentions.* The central part is section 7 *A Proposal* where I present my view that in using language (i.e. speaking) and consequently also in language change, we do not need to help ourselves with intentions. It is the claim that in speaking we do not have to form a thought about language, i. e., we do not have to form an intention when speaking. Consequently this is also true for language change. The strong claim is that speakers' having intentions do not have an explanatory role in language use or language change. If this is true then a futher tentative suggestion is that if the locus of change is not the individual mind (individual intentions), then the driving forces behind language change are/might be social. The intentions might have a role to play

intentionality as a property of doing things in a way that distinguishes humans from the animal world.

on higher levels, that is, in speakers' communicative events/attempts. This posibility is further explored in section 8 under the subtitle *Goals of communication* where I argue for the levels of explanation in language change. Section 9 briefly introduces the emergentist approach in linguistics as another possible theoretical framework for explaining language change and the short attempt is to relate it to the emergentist approach in the explanation of biological evolution as suggested by Denis Noble. In 10 *Concluding remarks,* I relate some of the highlights of this paper to Kathy Wilkes's comments on language and intentions in language.

2. *Causes of linguistic change*

Historical linguists, and linguists in general have always concerned with the question of why languages change. However, most of the explanations and answers provided in the past have been rather fanciful. Jespersen (1922), for example, enumerates a number of them, starting from anatomical reasons ("sound changes must have their cause in changes in the anatomical structure of the articulating organs" (255), then geographical ("the harsh consonants found in the languages of the Caucasus as contrasted with the pleasanter sounds in regions more favoured by nature" (256), to psychological ("since the times of Grimm it has been usual to ascribe the well-known consonant shift to psychological traits believed to be characteristic of the Germans… their progressive tendency and desire of liberty" (258). One of the most popular reasons given for language change was also the breathing efforts in mountain environment. Less outlandish reasons are given, as the imperfect language transmission, ease of articulation, laziness theory, etc. Jean Aitchison says "when we have eliminated the 'lunatic fringe' theories, we are (still) left with an enormous number of possible causes to take into consideration" (1981: 112).

Focusing on current literature, let us look at two models of language where we find reasons given for the explanation of language change. In one of these models, language is seen as an autonomous system (predating the birth of sociolinguistics in the 1960s) where speakers do not play any role in changing the language. In the other model, the so-called "rational agent" model of language, speakers play a role in language change. In this model speakers' intentions become important.

3. *What kind of "beast" is language?*

3.1. *Language as an autonomous system*

Before the 1960s with the birth of sociolinguistics, there was little or no systematic study of the possible roles of speakers (in social interaction) as initiators or carriers of change. The language-internal position was the default position in the explanation of language change (with a relative neglect of contact phenomena). American historical linguist, Roger

Lass, is a good representative of the model of language approached as an autonomous system. His ideas are presented in Lass (1980) and elaborated in Lass (1997).

Lass does not believe "that language change is the result of 'human action' except in a very distant, secondary and probably uninteresting way" (1997: 337). Lass is very supportive of Sapir's idea of language drift. The analysis of a drift, says Sapir, is certain "to be unconscious, or rather unknown, to the normal speaker" (1920: 161). If this is the case then, for Lass, language change cannot be a speaker's "act". (1997: 367). Lass believes that one has to include "the 'geological time' dimension, where speakers are not conscious of their role in propagating variation, and indeed can't be… just because a person happens to do something, this is not necessarily an 'act' (in the sense of representing a cognitive choice or anything of significance to the person). One can act out of tradition, habit, uncontrollable impulse (endogenous or drug-induced) or for no apparent 'reason' at all" (1997: 374).

Lass sees language as "a population of variants moving through time, and subject to selection" (1997: 377). His arguments, he believes, "point the way towards a reasonable, non-individual and non-social definition of what we mean by 'a language'" (1997: 375), where speakers' role in language change are totally excluded.[2] "In this view, language change was seen, like geological change, to be the result of powerful non-human forces, in which human goals and actions had no part" (1997: 387). For the "rational agent" model (to be presented next) Lass says: "The fundamental error of the hermeneutic approach is that it attempts to get 'inside' something that because of its immense historical extension may not have an inside at all (1997: 390).[3] What I have been trying to do…has been not much more than an attempt to get away from viscera and projections and pseudo-causal mysticism into something more like fresh air" (1997: 390). To sum up, it was believed that change in language is change in linguistic systems, not change in the speakers. Speakers are seen as powerless and insignificant figures.

3.2. *The "Rational agent" model of language*

The "rational agent" model of language is well represented by James Milroy (2003), especially because he goes into open discussion/dispute with Lass. Milroy's position is in great opposition to Lass. The hypothesis that language is a kind of abstract object that can change within

[2] "By saying we don't 'need' speakers I am not of course making the *absurd claim that language change proceeds entirely in their absence"* (Lass 1997: 377, note 42).

[3] "This dichotomy [between autonomous and agent centered] has been noted before, perhaps most perspicuously by Raimo Anttila (1992); it focusses particularly on that style of linguistic enquiry that rejects hermeneutics and/or neo-Aristotelian 'finalism' vs. the one that embraces it. Other names for the dichotomy might be 'classical' vs. 'romantic', 'sceptical' vs. 'enthusiastic', even perhaps 'rationalist' vs. 'irrationalist', 'agnostic' vs. 'missionary', 'Apollonian' vs. 'Dionysian'" (Lass 1997: 389).

itself or perhaps *bring about* change within itself, is a general nineteenth-century view and Milroy seen Roger Lass as a prominent, but balanced, defender of this traditional view. Milroy says that Lass[4] has correctly pointed out that in the tradition, it has been assumed that it is languages that change and not (necessarily) speakers who change languages[5] and that "endogenous change is part of the nature of the beast" (Lass 1997: 208). What is important is that the agency in this approach is language itself, and not the speakers of the language.

For the autonomous view of language change Milroy says a bit ironically: "'Good heavens!', says the language, 'I'm becoming ambiguous. I'd better use my prepositions to make myself clearer!'" (2003: 151). Milroy argues (as all of the sociolinguists do) that speakers/listeners play a vital role in language change, and that in addition, language changes in response to changes in external (social) conditions (2003: 146). Sociolinguistic or rational-agent model thus makes a necessary contribution to explaining language change *via* the role of the speakers. The promise is that sociolinguistic approach may help us understand how language systems move from one state to another due to the role or intervention of the speakers and social environment.

Milroy-Lass dispute is very interesting in its own right. But why is the above opposition to language and language change important for our discussion? If the model of language is speaker-based, then the role of intentions and speakers' actions in the explanation of language change becomes quite central. Are speakers doing something intentionally to language, do they deliberately set out to bring about changes in language? General agreement, however, is that speakers do not change their language with the aim of changing the language. Thus, Milroy approves of Lass pointing to "the implausibility of the view that *speakers* take action to *prevent,* for example, 'dysfunctional' changes" (1997: 359). Speakers do not care about the language in that way and moreover, we do not see into their minds. If the above is true, then what are speakers doing, what kind of actions should we ascribe to them? Before proceeding let us look into the most recent approaches to language change modeled on the evolutionary theory.

4. *Transferring the evolutionary metaphor language and language change*

The transfer of ideas from biological evolution to language is not a new one. The close relationship between biological evolution and language was noted by Darwin himself in an oft-quoted passage from *The Descent of Man*: "The formation of different languages and of different species, and the proofs that both have been developed through a gradual process, are curiously parallel" (Darwin 1882). During the last few

[4] Lass (1980: 120).

[5] Milroy (2003: 143).

decades it has become fashionable in linguistics—and in some other human sciences—to look to the theory of evolution for a new explanatory framework.[6] A number of books appeared transferring the biological metaphor to language and language change, the most important being Keller (1994), Saliloko (2001), Croft (2000), Givón (2002).[7]

I will discuss in broad outline William Croft's book *Explaining Language Change: An Evolutionary Approach* (2002). The main reason is that since our topic are intentions in language change Croft discusses them more than others do. Croft's approach assumes a usage-based evolutionary model, i.e., language change occurs in language use. Furthermore, variation in language is a crucial factor in language change. The background belief is that there is profound relationship between biological evolution and language change. Croft takes David Hull's application of evolutionary theory to conceptual change. Hull's conceptual system is referred to as the generalised analysis of selection.[8] Simply put, Croft adopts and adapts the theory of biological evolution in order to construct an evolutionary theory of language change. Language change is an example of the same process, or a similar process as evolution, occurring with a different type of entity, namely language. He tries to show that mechanisms and processes that are postulated by evolutionary theory in biology can be applied to language change. The evolutionary framework requires that the object of the study be a historical entity, i.e. a spatio-temporally bounded token, not an idealised natural kind. In language change, the paradigm interactor is the speaker, or to be exact, the speaker's grammar. *The only real place for a linguistic system to reside is in speaker's head.*[9]

[6] For example, the writings by Richard Dawkins (1986), Daniel Dennett (1995), David Hull (1988), and Gary Cziko (1995). Anette Rosenbach (2008) has a very thoughtful review of the problems and successes of such views and approaches to language.

[7] Mufwene (2001) also invokes evolutionary theory in his approach to language change. He calls language a parasitic species, because languages can only exist through their hosts, i.e. speakers. Ritt (2004) on the other hand, although supporting and advocating a Darwinian approach to language change sees speakers as "victims" of language change rather than agents.

[8] See David Hull (1988). In this work one of Hull's concerns is to define an evolutionary process in a way that could be applicable both to biological evolution and to the development and spread of scientific ideas.

[9] Here are some basic concepts into which we cannot go in this paper. The counterpart of DNA in biological systems is the utterance in language. *Utterance* is a particular, actual occurrence of the product of human behaviour in communicative interaction. *Language* is defined as the population of utterances in a speech community, the set of actual utterances produced and comprehended in a particular speech community and *Grammar* is the cognitive structure in a speakers' mind that contains their knowledge about their language, the structure that is used in producing and comprehending utterances. In gene-based biological selection, perpetuation of replicators, i.e. genes, is achieved by reproduction by the interactor, i.e. the organism. Reproduction may result in altered replication of the gene. In language change we have a *replicator* which is an entity that passes on its structure

Evolution is a two-step process: there is altered replication of the replicators (innovation), and then selection. The causal mechanism of evolution in language change is also a two-step process: there is i*nnovation* and then *propagation.*

In *altered replication* or *innovation*, the outcome is different in structure from the original (e.g. *bad* may be pronounced with a slightly higher vowel than one heard before). *Selection or propagation* is a process of perpetuation of relevant innovations in a community of speakers.[10] The emergence of new variants is treated differently from their spread through a speech community. In biology, the novelty emerges from the blind recombination and mutation of DNA.[11] The question then appears to be: Is the innovation in language also random or not? Opinions differ. Under one view variation in language arises randomly, like variation in biology and it is only the process of selection which brings in "order" into language change (McMahon 1994: 337). On the other view, variation arise non-randomly as, for example, argued by Haspelmath (1999: 192). He says: "I argue against the view that the grammatical constraints could be due to accident" (1999: 180). If errors in linguistic replication are in the same way random and non-optimizing as are errors in DNA replications, then it has consequences for the innovation of a linguistic variable and for the role of speakers' intentions in the innovation of a new variants. There is more uniformity of opinions about the selection process. Croft (2000) for example argues that it is social factors—and *only* social factors—that drive the selection process. He refers to the main determinants of linguistic choices known from the sociolinguistic literature, such as accommodation (adaptation of one's speech to that of an interlocutor) prestige (overt and covert), relation to social parameters as class, gender, age, etc.

Evolutionary approach to language changes underwent a number of criticisms. Let me just mention some by Andersen 2006.[12] Andersen claims: 1. That an innovative reanalysis in language is not random but that it is recognizably rational. 2. That there is nothing in the replication of genetic material that corresponds to the *imposition of values* on content and expression elements which takes place in the process

largely intact in successive replications. *Interactor* is an entity that interacts as a cohesive whole with its environment in such a way that this interaction *causes* replication to be differential. Differential replication is an innovation in language system.

[10] The stress on variability in language and the distinction between actuation/innovation and selection/propagation is essential in this theoretical framework. It has been so since the pioneering article by Weinreich, Labov and Herzog (1968) in their famous statement that "[n]ot all variability and heterogeneity in language structure involves change; but all change involves variability and heterogeneity" (1968: 188).

[11] Cziko says: "Darwinian mechanism of cumulative blind variation and selection is the only tenable nonmiraculous explanation for the emergence of any kind of functional complexity" (1995: 300).

[12] See the exchange between Croft and Andersen in Nedergaard Thomsen (2006).

of reanalysis. 3. In actualization speakers "*literally select some variants over others*" (italics mine) but in natural selection there is never any agent purposefully producing an action of selecting something over something else. In other words, in evolution there is blind mutation natural selection while in language change we have rational speakers who make choices. Andersen says that the statement "Danish has adapted to the computer age" is really short for the equivalent that Danish speakers have innovated, adopted, and integrated (a linguistic feature) into their tradition of speaking. In sum, the mechanical replication of genetic material in evolution contrasts with the rational process of reanalysis in language change. "If so, then here is a sharp contrast between evolution and language history: while genetic copying errors result from an underperformance of the mechanisms of replication, the formation of grammar (and other cultural systems) demonstrates an overperformance of human minds, a capacity for forming new symbols for immediate use that surpasses any need to acquire precisely all the details of extant patterns of usage" (2006: 81).

Andersen believes that change in language is produced by its speakers as part of the exercise of their *free will* which, according to him, speaking is. Speakers as free agents (with their human minds) are the agents of change. When one is dealing with structural and developmental tendencies in language it is in the linguistic behavior of speakers that is most important. So why does language change, according to Andersen? Apart from the already mentioned free will, it is "the *creative aspects of practices* and traditions of speaking that matter. The fact that they leap to the eye in every type of innovation that has been described suggests that they are not an accidental, but an essential characteristic of language" (2006: 83).

If speakers are free and creative agents and they are the locus of language innovations as it is claimed in the "rational agent" approach to language and language change, then the talk of intentions becomes very relevant or crucial in the explanation of language change. James and Lesley Milroy (1985) follow the same line of thought. Change begins with variation in the speech of speakers. They affirm that if we are to address the actuation problem (which is "the very heart of the matter"), we must break with tradition and maintain that it is not languages that innovate. It is the speakers who innovate and their role is essential. In the evolutionary model of language change which is supposedly mechanical and blind one would expect that the role of speakers is minimized or non-existent. But this is not the case. On the contrary, speakers are, in this model also, central for the explanation of change which sounds controversial or even contradictory—if the change is blind and random. We examine what has been said about speakers' intentions in the next section.

5. *On speakers' intentions*

What has been said about speakers' intentions? Apart from some scattered remarks and the stress on conscious or problematically unconscious intentions by the speakers, there is no systematic approach to their discussion in the past. Here are a few examples. Whitney (1848–1916), for example, held the view "that language change is governed by two different forces—conscious intentional action (individual variation) and 'unconscious' consequences (social selection)" (Nerlich 1990: 40). Bréal (1832–1915) thought that the language user is the motor of change, that language change is the cumulative consequence of intentional, intelligent, and conscious actions of the speaker. Language change "has to be explained by reference to conscious, voluntary action (Nerlich 1990: 104). Changes are bought about unconsciously, however "by an *unconscious that has depth*, so to speak... consciousness plays a role in language" (Nerlich 1990: 104, italics mine).

What do we find in the authors that were discussed so far, namely, Andersen and Milroy in particular, since they put speaker at a center position for the explanation of language change?

We have already seen that Andersen sees the speaker as a rational agent imposing values on content and expressions and "doing something" in the course of linguistic change. If this is so, one would expect that intentions will be discussed in great details. But then Andersen expresses his doubts about intentionality in a longer passage that I quote:

> But such a reference to intentionality is inappropriate for several reasons. For one thing, we rarely know much about the intentions of the speaker(s) that initiated or adopted past innovations. For another, there are evidently several kinds of intentionality. Experience tells us that Adaptive innovations and Extensions can be created with premeditation—consider the Coining of new terminology or metaphoric Extensions in poetry. If Adaptive innovations and Extensions are not premeditated, they can still be made deliberately. But even if an innovation is not made deliberately, but spontaneously and seemingly unwittingly, the speaker may still be able to rationalize it afterwards, that is, it may appear to have been made with unconscious intent. *This fuzziness of the notion of intention speaks in favor of shifting our attention from the innovating speaker's inscrutable state of mind to the purpose or purposes served by given innovations*: all Adaptive innovations and Extensions are purposeful (2006: 68, italics mine).

What has to be noticed in this passage in particular is that Andersen in his hesitancy to speak of speakers' intentions switching the explanatory aim to *the purposes of communication.*

James Milroy mentions speakers' intentions under the subtitle "Intentionality and change" in his article from 2006. All he says is: "It does not follow from speaker-based position arguments that speakers deliberately set out to bring about change in language…we do not see into their minds…they care (not) about the language…Although speakers do not voluntarily engineer changes, it must be speakers who

implement them in ante action and who finally determine, through frequency of use, which changes, out of a very large array of possible changes, are accepted into the system" (2006: 149–150). One can surely interpret that Milroy does not think that in language change speakers "having intentions" (no. 2) play any role. And he is surely right as I shall argue later.

What do we find on intentions where language is approached from the evolutionary model as applied to language change? If the evolution is blind then by analogy language change is blind, it is a result of chance, it is random. So, what is the role of the individual (and his intentions) in the evolutionary based approaches? One would expect that the stress on the individual role in language change should be minimal. But this is not so. On the contrary, the attempts are to show that the individual and his/her intentions are still very central and quite prevalent. Rudi Keller (1990) spends a number of pages on the role of intentions.[13] For example he says that "'the speakers change their language' only sounds inappropriate because the speakers do not change their language intentionally and systematically but unconsciously" (1990: 8–11). He questions the status of conscious *vs*. unconscious intentions and does not support the claim that unconscious intentions are problematic. Keller sees language change as what he calls a phenomenon of the third kind, i.e., an unintended causal effect of intended human social actions (1990: 57). The phenomenon is said to be of the third kind to distinguish it from the products of intentional design (artifactual phenomena) and products of purely natural processes with no involvement of human intentions (natural phenomena).[14] Language change is the causal consequence of a multitude of intentional actions. Thus, individual intentional actions (unconscious?) are involved in language change (1990: 68). At other places Keller is more outspoken and says: However, "conscious human purpose is always involved" (1990: 86). Furthermore there is no crucial influence on language, without going through the freedom and the intelligence (?) of the speakers (1990: 90). In sum, "there is always a conscious purpose involved, as in any communicative activity, whereas change is its (usually) unintended cumulative effect" (1990: 121). Languages do not change in certain ways because speakers intend them to do so, but they change as a by-product of the speakers' intentions to attain socio-communicative goals with their language use. We shall comment on these claims in the next section.

[13] He points out the ambiguity and different meanings of intentions. He finds the problem of terminological confusion of lumping three terms: intentional, planned, and conscious together. Intentional is sometimes confused with planned but these are predicates which are independent of each other. Here is an example: "When I am about to open the door, I moved the thumb from the index finger to grasp the door handle. This action undoubtedly has a purpose. It is goal-directed, but I never planned to do it" (1990: 10).

[14] As mentioned in Croft (2000: 59–62).

Let us return to William Croft (2000).[15] Croft, as we saw, warns against the "reification or hypostatization of languages...languages don't change; people change language through their actions" (2000: 4). What we find in Croft and not in other linguists who talk about intentions is an attempt to systematize speakers' intentions into: nonintentional and intentional.

Here is a relevant part of the chart:

	Intentional	*Nonintentional*
Normal replication	convention (being understood)	entrenchment
Altered replication (innovation)	expressiveness not being misunderstood economy	over/undershoot (hypercorrection hypocorrection) form-function reanalysis [speech errors]

In *normal replication* the *nonintentional* mechanisms are found in entrenchment. What Croft means by entrenchment is the psychological routinization of a behavior, i.e., the behavior of recognizing a linguistic expression and producing it (2000: 236). The entrenchment is the survival of the cognitive structures in a grammar that are used by the speaker in producing utterances of that structure. On the other hand, Croft finds *intentional mechanisms* in language convention which is a common ground in a community. I will later question this decision.

Let us look at the suggestions for altered replication, that is innovation:

Nonintentional mechanisms for innovation are: speech errors, sound changes, hypercorrection and hypocorrection. Croft says: "the speaker aims to produce a particular sound, but overshoots or undershoots the target …" (Croft, 2000: 76).[16]

[15] There is no space to discuss Ritt (2004) but it is interesting to see the difference between Ritt's and Croft's approach concerning the role of the speaker in linguistic replication. Ritt (2004) adopts Dawkins's (1976) notion of "selfish genes" and thus Dawkins's idea that memes actively replicate and that the organism's (i.e. the speaker's) role is simply that of a "vehicle", i.e. speaker has a very passive role. Croft (2000), in contrast, adopts Hull's generalized theory of selection and with it Hull's idea of somewhat more active "interactors" rather than Dawkins's passive notion of "vehicle".

[16] An example of hypercorrection would be: *It is I*, or *seldomly* and of hypocorrection the nasalization of *can* (*kan*). An example of form-function analysis would be: *He robbed her of her bracelet* as differently expressed: *He robbed the bracelet from her* showing the flexibility of recombining existing forms-cum-meanings.

Intentional mechanisms of innovation are: expressiveness (creativity), avoiding misunderstanding, and economy. Croft says that one of the chief mechanisms for innovation in lexical change is *the slipperiness of meaning.*

6. *Problems with explaining language change with speakers' intentions*

What possible conclusions can we draw from the writings on intentions as playing a role in language change?

1. One finds the discussion controversial and insufficient to say the least.[17] 2. More specifically, a number of claims on the role of intentions are contradictory. Andersen, for example, says that adaptive innovations (like coinage or borrowing) may be considered intentional, and extensions (application of extant means to new usage, received lexeme to a new referent), unintentional. But later he expresses his doubts and says that if adaptive innovations and extensions are not premeditated, they can still be made deliberately. If they are made deliberately then they cannot be nonintentional. Keller stresses that language change is a causal consequence of a multitude of intentional actions. But then he also says that languages do not change because speakers intend them to do so. So speakers change the language intentionally but then it seems that they do this unconsciously. In other words, Keller allows for unconscious intentionality. He also talks about the power of "free will and necessity" as a cause of language change, which, he claims, should corresponds to the interaction of the factors like "chance and necessity" in the evolution of animate nature. Frequently people assume that chance allows for free will, while in fact it is difficult to see how random, chancy phenomena allow for free will. 3. Thirdly and possibly most importantly, when linguists are using intentional it is not clear if intentional is used as "doing A intentionally" (1) or it is used as "having intentions to do A" (2). In his hierarchical view of intentions Croft says: "Certainly normal replication—adherence to convention—is an intentional mechanism that nonintentional mechanisms cannot do without" (2000: 78). Yes, if by intentional mechanism Croft means intentional actions (1). No, if it means that in conventional, normal/everyday language use we as speakers help ourselves with having intentions (2). Having intentions (2) do not have a place in the explanation of language conventions. At least, I want to argue for this view in the next section.

7. *A proposal*

Intentions used in the explanation of language and language change seem to have a number of problems: 1. Unanswered questions (what is

[17] Looking at the indexes of many books on language change we find very few entries, if at all, on speakers' intentions.

unconscious intention?), 2. Confusions (intentional actions vs. having intentions), 3. Contradictions (free will vs. blind selection).

A good methodological strategy is to seek nonintentional mechanism first, and only turn to intentional mechanism at higher linguistic levels. The reasonable suggestion is that nonintentional mechanisms for innovation are more likely to be found at lower levels of language organization such as sound structure, while intentional mechanisms are more likely to be found at higher linguistic levels (Croft 2002: 76). In this respect Croft's hierarchical structure as presented above is useful as a starting point.

In a possible hierarchical structure, I first follow John Ohala (1989) whose research is mainly in phonology and who also deals with issues of phonological change. Ohala is a firm advocate of the elimination of intentional talk on the phonological level. For him the source of variation is definitely the speaker but the speaker is *unaware* of the variation. He says: "There exists in any speech community at any point in time a great deal of hidden variation in the pronunciation of words.... by hidden I mean rather that speakers exhibit variations in their pronunciation which *they and listeners usually do not recognize as variation*" (1989: 175, italics mine).[18] Speaker is totally unaware of any kind of change so like in biological evolutionary theory "there is no mind directing the change, no choices made to take one path over another" (1989: 33). Ohala justifies the exclusion of speakers' intervention, i.e. speakers' intentions, in language change with a somewhat unusual comparison and he says: "I avoid explanation of the sort '...the speaker chose a different pronunciation in order to optimize (something)'...for the same reason that modern science rejects explanations like '...the earth's climate is getting warmer because the gods are angry with us'.... it is part of the tradition of modern science to seek the less extravagant explanation before embracing the extravagant ones. This is, after all, the nature of explanation: reducing the unknown to the known ...not to further unknown, uncertain, or unprovable entities" (1989: 37). It is obvious that Ohala finds the intentional talk in language change nothing more than an extravagant myth not worthy of being part of a scientific approach to language.

Ohala is concerned only with the initiation (actuation) of sound changes, not their transmission. The way that change gets transmitted is by ordinary means of reproduction (1989: 21). "Spread is mediated primarily by psychological and social factors and lies outside the domain I consider here" (1989:15). In innovation Ohala is looking and supporting mechanistic or nonintentional causes of change. In sum, the claim is that there is no need, and moreover it is implausible and scientifically wrong, to include speakers' intentions in phonological change.

[18] And more strongly: "What I am claiming is that the devoicing of voiced stops and the frication of stop releases can happen inadvertently or unintentionally" (1989: 178). Ohala takes sound change in its initiation (or innovation) to be nonmentalistic.

What I want is to suggest (more radically) that normal language use and with that language change is not intentional at all. In other words, we do not need intentional talk in order to explain ordinary language use or language change on any linguistic level by invoking speakers' intentions. My suggestion is that Ohala-style explanation should be extended to higher linguistic levels such as morphology, syntax and even lexicon. In other words, in everyday language use and with that language change there is no need to invoke speakers' having intentions at all. They do not play any explanatory role in conventional language use. The speaker is not intentionally doing anything (2). He is only intentionally acting (1). But this is as it should be.[19]

In order to accept this proposal we have to take for granted some background theoretical assumptions. 1. We have to see linguistic competence not as knowledge-that (even tacit) but as a skill or ability, i.e., knowledge-how. I go along with Devitt here who says: "Why think that linguistic competence is just a skill or ability? Briefly, because it has all the marks of one: it has limited plasticity; it is extraordinarily fast; the process of exercising it is unavailable to consciousness; once established, it is "'automatic' with the result that it can be performed whilst attention is elsewhere" (2020: 28).[20] 2. Furthermore, one has to accept that conventions play a significant role in language. A convention is the regular use of language forms on all linguistic levels and speakers in a community are participating in the same (or very similar) linguistic conventions. Devitt says: "These shared dispositions amount to a linguistic convention if their sharing is explained by a certain sort of causal relation between the dispositions" (2021b: 83). Regularity is noticed by speakers and hearers but (very importantly) "this noticing and catching on are likely not high-level-cognitive processes; likely, they are 'implicit' and 'procedural' rather than 'explicit' and 'declarative'" (2021b: 86).[21] If following the conventions in language use is not high-level cognitive process then speakers do not have to use intentions in order to speak or change their language. This is why I think that Croft is not right in putting convention (being understood) as an intentional mechanism.[22]

If the above is accepted (and I do not claim that it is not controversial) then were do we find having intentions as playing a role in language change? In the hierarchical structure where can we find place for intentions? If asked what kinds of linguistic changes speakers are most

[19] To be reminded of the comparison: We walk intentionally but we do not form an intention to walk.

[20] See also Devitt (2006b: 209–10). I argued for knowledge of language as knowledge-how and not implicit or tacit knowledge-that in Jutronić (1995).

[21] See also Devitt (2006b: 210–20).

[22] Devitt in his article "The irrelevance of intentions to refer" has argued convincingly that reference fixing does not need any use of intentions, either (2021b). He finds the explanation with intentions "implausible, incomplete, redundant once completed and finally misleading." Indeed, intending to refer "should have no place at all in a theory of language" (2021b).

likely to make deliberately, one would think first of lexical innovations. Possible conscious role of individual speakers is especially clear in lexical innovation cases of new words created by high prestige individuals, such as writers and poets. Every generation of teenagers has its own slang vocabulary and every specialized field has its own technical lexicon. There are words that invented either entirely (e.g., names of new products such as Kleenex and Xerox). Or partly to take an obvious example: email, for instance, combines the first letter electronic with the noun mail, etc. Metaphorical use of language is also intentional, not to mention poetic use of language. All the uses of language that pragmatists try to stress, those due to contextual factors and interpretations are likely to be intentional. A very important thing to notice is that in the above stated possibly intentional use of language and the inclusion of pragmatists' claims we are not talking anymore about ordinary language use. The talk has switched to communication, its strategies and its goals.[23]

8. *Goals of communication*

What do the authors we discussed say when trying to explain language change? Whichever approach is taken, either autonomous or agent driven or the approach on the model on evolutionary biology, when one looks more carefully one notices that in trying to explain the innovation in language the authors often, one might say, change the subject from individual actions to the goals of communication.

Croft states in the above chart that intentional mechanisms for innovation (his altered replication) are: expressiveness (creativity), not being misunderstood, economy. In Andersen we find the stress on the rationality of the agent, his free will as evident in creation of new words and poetic language. What one notices is that the mentioned mechanisms have little to do with ordinary language use. In the proposed hierarchy of nonintentional and intentional mechanisms their place is to be found in the communicative strategies and not in language as a conventional means of communication. Expressiveness, creativity, not being misunderstood, economy, not to mention free will and rational choices are mechanisms not involved in ordinary, nonintentional language use and language change. I suggest that intentions have an explanatory role in what I labelled as goals of communication. The suggestion itself is actually nascent, although mostly covertly, in the writing of the authors involved in this discussion.

For example, in Milroy with his speaker oriented assumptions, we expect to hear more about speaker's intentions but when Milroy asks who practices bricolage in language he switched from the role of the speaker to the discussion of speaker's *communicative strategies* (2003:

[23] See for example Devitt (2021a) for the critical debate about pragmatists' claims and where to draw the line (distinction) between semantics and pragmatics.

156). He says: "…the change that I am about to discuss here is involved with the *communicative strategies* of speakers" (2006: 257); "no change is ever independent of some form of *speaker-based social motivation"* (2006: 161). Keller argues that the linguistic change is a by-product of the speakers' intentions to *attain socio-communicative g*oals with their language use. Croft makes very much about the distinction of speakers' innovation and selection (or propagation) which he says is intentional. "*Language use* is intentional behavior. What matters, however, is *the goal of the intention*" (2006: 119, italics mine). Lass who emphasizes the implausibility of the view that speakers take action in language change says: "… they [speakers] are preeminently interested in *communication*, and do not deliberately and consciously aim at changing language" (1997: 359). Isa Itkonen who (like Andersen) sees language change as rational action of human free will, reverts to a community of speakers and gives them an important role in the selection of certain innovations. "The real effective reason of a given (phonetic) change is that a *community*, which might have chosen otherwise, *willed it to be thus*..." (2005: 73, italics mine).

What can we reasonably conclude from the above statements or claims? One thing seems to be certain and that is that the discussion of speakers' intentions in language change is switched above linguistic levels, to the level of communicative interaction with the stress on the goals of communication. They all support the sociolinguistic guiding idea that the most significant contribution of sociolinguistics to linguistics in general is the fact that is has been demonstrated time and again that one cannot fully understand the emergence, spread and loss of a linguistic feature without taking into account extralinguistic factors. As Labov, the father of sociolinguistics says: "rarely do we have some sense of what gets the whole thing rolling in the first place in terms the 'actuation problem'" (1972: 162–63). "Therefore we can say that the language has changed only when a *group* of speakers use a different pattern to communicate with each other… The origin of a change *is* its 'propagation' or acceptance by others" (Labov 1972: 277). However, there is also a general conviction that processes of linguistic change are "multi-causality" phenomena in the way that *cognition and social structure interact* and shape the path of language change. But maybe one has prevalence over the other in the role they play in the explanation of language change. In a larger perspective set forth by Weinreich, Labov, and Herzog (1968), we can say that the linguistic behavior of individuals cannot be understood without knowledge of the communities that they belong to. They give prevalence to social factors. All the observations Labov made in Martha's Vineyard gave him the idea that speech is always linked to social attitudes and linguistic change of several groups of society.

If the stress in language change is switched from speakers to their goals in communication, then do we have to switch from the individual

to the collective? Peter Harder (2010), for example argues that the individual is a wrong starting point in approaching language and thus also language change. His contribution is in suggesting how cognitive linguistics is to be expanded to include the social side of language and meaning. In other words, language-and-conceptualization needs to be set in the wider context of "meaning-in-society". Language and language change are fundamentally social interactional phenomena. "If a word meaning does not exist in a sociocultural niche (however fleeting and emergent), the word does not exist at all" (2010: 171). But "if we see the existence of meaning at collective level…, the fact that meaning cannot exist without individual minds is no argument against collective meaning" (2010: 166).

One of the more important goals in communication that one finds discussed in literature is speakers' attempts to accommodate to their interlocutors. This is known as language accommodation. Very briefly, research shows that in the process of accommodation speakers tend to adapt/accommodate their language to the interlocutor, which necessarily gives rise to linguistic change.[24] Communication Accommodation Theory (CAT) shows that interlocutors tend to converge linguistically over the course of interaction (Giles 1980). The goal of accommodation is possibly an intentional mechanism for language change. There is (intentional) convergence in face-to-face interaction. For example, in contexts of dialect contact speakers accommodate their variety to other variety or varieties in order to show solidarity, identity, etc. The variants that emerge are a result of accommodatory behavior which gives rise to linguistic change and which can/may gradually stabilize and become more durable characteristic of that person's linguistic repertoire.[25]

In sum, in weighing the role of individual/mental and social we might conclude that cognitive states have to be completed with a reflective social evaluation. There surely are different unreflective, non-intentional cognitive/perceptual factors that contribute to innovation but again if they are not completed with reflective, intentional social evaluations, we would not surface at all, i.e. we would not know about them at all.[26]

9. *Emergentism*

A possible more theoretically profitable way to look at hierarchical levels of intentional talk is within the emergentism approach. Emergentism in linguistics is becoming more and more popular. The advo-

[24] See for example Trudgill (1986).

[25] See Kerwill (2002). On the other hand there are opposing views to intentional explanation of accommodation. For example, Trudgill says: "linguistic accommodation is not driven by social factors such as identity at all but is an automatic consequence of interaction" (2008: 252).

[26] See Jutronić (1995).

cates of emergentism characterize both the language of the community and that of the individual as being in a state of *constant change* and reorganization. The idea is related to the explanation in usage-based linguistics in emphasizing that language structure emerges from language use. Linguistic emergentism assumes that the properties of language arise from the interaction between the demands of communication and general human capabilities. The issues are numerous as evident from the articles in the recently published volume *The Handbook of Language Emergence* in 2015 that has over 600 pages. The core idea uniting this approach is that levels of linguistic structure emerge from patterns of usage across time. It firmly embraces the idea of inherent variability and uses variationist (sociolinguistic) tools for tackling specific descriptions and problems. There is a lot of stress on an interlocking hierarchical structure that is of interest to us here. Complexity arises from the hierarchical recombination of small parts into larger structures. Given the interactive nature of these interlocking hierarchies, reductionism (Fodor 1983) is clearly impossible. Within the emergentist framework, the principles of competition, hierarchicality, and timeframes are recognized and much discussed.

In their contribution on linguistic change in the emergentist framework, Poplack and Cacoullos (chapter 12), trace changes and continuities in grammar and lexicon over decades and even centuries. They view the individual's linguistic abilities as emerging from interactions with the wider social community. They refer to sociolinguistics as "language emergence on the ground" because of the richness of its observational data relating to language usage and change. They show that by situating newly emerging forms in the social and linguistic structures, we can discover the mechanisms involved in emergence of new forms. A core insight of this approach to language is that form–function mappings are inherently variable and there is mention of Darwinian theory in producing and proliferation of variants.

Is the emergentism approach another possible venue of discussing the role of speakers' intentions in the explanation of linguistic change? I think that the answer has to be: No. One notices that emergentists hardly mention intentions at all. The index of the mentioned volume barely has an entry or two on intentions or intentionality. Thus, even simply looking at the index, one will conclude that authors do not seem to be interested in intentions in language or language change. The whole stress again is that language changes across generations is hierarchical manner and that the changes are determined by communicative function.

Denis Noble in his book *The Music of Life: Biology Beyond Genes* (2006) argues for (if I read him right) a hierarchical multilevel selection view in biological explanation which is not gene-centered. He is proposing an emergentism view of higher-level properties. He says: "This, then, is the great challenge of twenty-first-century biology: how

to account for the phenotype in terms of the systems-level interactions of the proteins" (2006: 17). Some biologists have called these properties "emergent" properties. Noble prefers to call them "systems-level" properties. The higher-level properties emerge from the lower ones and linking levels is part of what systems biology is about (78). One of the important goals of integrative systems biology is to identify the levels at which the various functions exist and operate (129).

For the purpose of our discussion it is important what Noble says about who is driving or creating the emergent properties. The parallel question for language is of who is driving or creating language change. And Noble's answer is: Nobody! He says: "'I' am nowhere to be found. The subject is not usually there.[27] It all has to emerge without there being a driver. The grand composer was even more blind than Beethoven was deaf!" (112). In our case, it is not the individual or his/her intentions that changes language. What is also important in the emergentist's framework is that "explanation is possible only at the appropriate level, in this case the level at which it makes sense to talk about…"(129). In the case of intentions in language, levels are important, too. The proposal was that the level where we can talk about intentions is in communication strategies and not in ordinary language use. Noble also stresses the importance of social context. "Obviously, any explanation of my pointing *as an action* would need to take that social context into account" (127). The same in language case. My suggestion was that the levels where we can talk about intentions is much more on the creative use of language and communicative strategies than in ordinary language use.

Thus, it seems that in our journey about the role of intentions as an explanatory tool in language change we have come back a full circle to Roger Lass who says: "… we don't gain anything by invoking them [speakers] (whatever their role)" (1980: 377, note 42). "There is of course no doubt that at some point in the procedure humans do have a role to play (individually and collectively), since they are at least end-users. The important thing is not to confuse the end-user with the product" (1980: 385).

10. *Concluding remarks*

We have gone a long way from presenting language as an autonomous system where a linguistic change is discussed as purely a language-internal account and external influences are not taken into account. Speakers' role in changing the language is minimized. Then subsequently speaker-based account of language change was found more satisfying but also more demanding. Speakers are agents, they bring about the language change. The role of speakers' intentions becomes

[27] "The most natural way of saying the Japanese or Korean equivalent would be 'thinking, therefore being'" (2006: 140).

rather crucial. But still, the goal of speakers' intentions is not linguistic change. The common agreement among linguistics is as Croft puts it, "[s]peakers have many goals when they use language, but changing the linguistic system is not one of them" (2000: 70). With the evolutionary model of language change the problematic nature of intentions becomes more evident. If linguistic change as evolutionary process is blind and random, then speakers as agents become problematic.

I then looked into the arguments for the crucial role of speakers' intentions (either conscious or unconscious) in understanding language change and found them either incomplete or insufficient. I suggested that ordinary language usage and also language change cannot be explained by intentional language.

So, who or what is changing the language? If the individual is not a good starting point then I suggested (after Harder's ideas and many usage-based approaches to language and language change) looking into the goals of communication. The crucial factors enabling us to explain the phenomenon of "language change" have, accordingly, to be localised to the social nature of human beings. Social and communicative aspects of linguistic structures require a *communication-centred perspective*. One example that I briefly discussed was accommodation theory. The most important suggestion put forward was that of the hierarchical order in the explanation of language change—from nonintentional to intentional actions and finally to speaker's intentions. Speaker's intentions play a role at higher levels related to creative language use and communicative strategies. I (tentatively) introduced the most recent attempt in emergentist linguistics where it is assumed that the changes of language arise from the demands of communication. I tried to draw the parallel to the approach in biological emergentist in the explanation of the evolutionary change as suggested by Denis Noble. Higher-level properties emerge from the lower levels.

Since we are celebrating Kathy Wilkes let me conclude with some of her views. Kathy says: "whether or not goal-representations, or intentions, are essentially cited in the explanation of purposive behaviour. I think it is obvious that they are" (1989b: 205). Kathy Wilkes is here saying that intentions are essentially cited in the explanation of purposive behaviour. In other words intentionally doing A requires an intention to do A. I tried to show that it does not. In the next quote Kathy Wilkes says: "He shot Lincoln, he pulled a trigger, he crooked his index finger. There comes the point, low down in a hierarchy, when we want to reject all talk of 'intentions'; it is to put it mildly, odd to say that the concert pianist 'intended' to hit C-sharp when playing a fast prelude" (1989b: 208). What is suggested in this passage is that intentions are not needed at the lowest point in the hierarchy which is similar to my main proposal that language is simply a skill and there is no room for invoking speaker's intentions at this level. Moreover, Kathy Wilkes mentions implicit and explicit intentions (maybe unconscious and con-

scious?) and she says: "We thus find a sliding scale from the apparently clear cases of explicit (or explicitly stated) intentions to those that seem 'merely' implicit..." (1989a: 162). This comes close to the suggestion of hierarchical order of levels of explanation with explicit intentions being introduced at higher levels that have to do with communicative strategies.

Needless to say the devil is in details about which, sorry to say, I have not said much.

References

Aitchison, J. 1981. *Language Change: Progress or Decay?* London: Fontana Paperbacks.

Andersen, H. 1973. "Abductive and deductive change." *Language* 49: 567–595.

Andersen, H. 1989. "Understanding linguistic innovations." In Breivik and Jahr 1989: 5–28.

Andersen, H. 2006. "Synchrony, diachrony, and evolution." In Nedergaard Thomsen (ed.). 2006: 49–91.

Anttila, R. 1992. "The return of philology to linguistics." In Pütz 1992: 313–335.

Blevins, J. 2004. *Evolutionary Phonology: The Emergence of Sound Patterns*. Cambridge: Cambridge University Press.

Bratman, M. 1984. "Two faces of intention." *Philosophical Review* 93: 375–405.

Breivik, L. E. and Jahr, E. H. (eds.). 1989. *Language Change: Contributions to the Study of Its Causes*. [Series: Trends in Linguistics, Studies and Monographs No. 43]. Berlin: Mouton de Gruyter.

Chambers, J. K., P. Trudgil, and N. Schilling-Estes, 2002. (eds.). *A Handbook of Language Variation and Change*. Oxford: Blackwell.

Cravens, T. D. (ed.). 2006. *Variation and Reconstruction: Currrent Issues in Linguistic Theory 268*. Amsterdam: John Benjamins.

Croft, W. 2000. *Explaining Language Change: An Evolutionary Approach*. Edinburg: Pearson Education.

Croft, W. 2002. "The Darwinization of Linguistics." *Selection* 3 (1): 75–91. http://www.akkrt.hu/journals/select

Croft, W. 2006. "The relevance of an evolutionary model to historical linguistics." In Nedergaard Thomsen 2006: 19–133.

Coseriu, Eugenio. [1952] 1975. "System, Norm und Rede". *Sprachtheorie und allgemeine Sprachwissenschaft*. Munich: Wilhelm Fink Verlag. Translated by Uwe Petersen from "Sistema, norma y habla." *Revista de la Facultad de Humanidades y Ciencias*, 9: 113–177, Montevideo, 1952.

Coussé, E and von Mengden, (eds.). 2014. *Usage-Based Approaches to Language Change*. Amsterdam: John Benjamins.

Cziko, G. 1995. *Without Miracles: Universal Selection Theory and the Second Darwinian Revolution*. The MIT Press: Bradford Book.

Darwin, C. 1871. *The Decent of Man and Selection in Relation to Sex*. London: John Murray.

De Silva, S. (ed.). 1980. *Aspects of linguistic behavior*. York: York Univer-

sity Press.
Devitt, M. 2006. *Ignorance of Language*. Oxford: Clarendon Press.
Devitt, M. 2021a. *Overlooking Conventions: The Trouble with Linguistic Pragmatism*. Cham: Springer.
Devitt, M. 2021b. "The irrelevance of intentions to refer: demonstratives and demonstrations." *Philosophical Studies*. https://doi.org/10.1007/s11098-021-01682-5
Dawkins, R. 1986. *The Blind Watchmaker: Why the Evidence of Evolution Reveals a Universe without Design*. London: Norton & Company, Inc.
Dennett, D. 1995. *Darwin's Dangerous Idea*. New York: Simon & Schuster Paperback.
Eckardt, R., Jäger, G. and Veens, T. (eds.). 2008. *Variation, Selection, Development, Probing the Evolutionary Model of Language Change*. Mouton: De Gruyter.
Eckardt, R. 2008. "Introduction." In Eckart, Regine et al. 2008: 1–23.
Fodor, J. 1983. *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge: MIT Press.
Giles, H. 1980. "Accommodation theory: Some new directions." In De Silva (ed.). 1980: 105–136.
Ginsburg, S. and Jablonka, E. 2019. *The evolution of the sensitive soul: Learning and the Origins of Consciousness*. Cambridge: MIT Press.
Givón, T. 2002. *Bio-linguistics*. The Santa Barbara lectures. Amsterdam: John Benjamins.
Harder, P. 2010. *Meaning in Mind and Society: A Functional Contribution to the Social Turn in Cognitive Linguistics*. Mouton: De Gruyter.
Haspelmath, M. 1999. "Optimality and diachronic adaptation." *Zeitschrift* für *Sprachwissenschaft* 18 (2): 180–205.
Hernández-Campoy, J. and Conde-Silvestre, M. (eds.). 2012. *The Handbook of Historical Sociolinguistics*. New York: Wiley-Blackwell.
Hickey, R. (ed.). 2003. *Motives for Language Change*. Cambridge: Cambridge University Press.
Hickey, R. 2012. "Internally and externally motivated language change." In Hernández-Campoy and Conde-Silvestre 2012: 387–407.
Hull, D. 1988. *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science*. Chicago: University of Chicago Press.
Ishiyama, O. 2014. "The nature of speaker creativity in linguistic innovation." In Cousséand von Mengden 2014: 147–169.
Itkonen, I. 2005. "The central role of normativity in language and linguistics." In Zlatev et al. 2005: 279–30.
Jespersen, O. 1922. *Language: Its Nature, Development and Origin*. London: Allen and Unwin.
Jutronić, D. 1995. "Knowledge of language." *Acta Analytica* 1995: 91–104.
Jutronić, D. 2015. "Cognitive Pragmatics and Variational Pragmatics: Possible Interaction?" *Croatian Journal of Philosophy* 15 (44): 233–247.
Kellermann, G. and Morrissey, M. D. (eds.). 1992. *Diachrony within Synchrony. Language History and Cognition: Papers from the International Symposium at the University of Duisburg, 26–28 March 1990*. Frankfurt a. M.: Peter Lang Verlag.

Keller, R. 1990. *On Language Change: The invisible hand in language*. London and New York: Routledge.

Koopman, F. W. et al. (eds.). 1986. *Explanation and Linguistic Change*. Amsterdam: John Benjamins.

Kerswill, P. 2002. Koineization and Accommodation. In Chambers et al. 2002: 669–702.

Labov, W. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Labov, W. 1994. *Principles of Linguistic Change, Volume 1: Internal Factors*. Oxford: Blackwell.

Labov, W. 2007. "Transmission and Diffusion." *Language* 83 (2): 344–387.

Lass, R. 1980. *On Explaining Language Change*. Cambridge: Cambridge University Press.

Lass, R. 1987. "Language, Speakers, History, Drift." In Koopmann et al. 1987: 151–177.

Lass, R. 1997. *Historical Linguistics and Language Change*. Cambridge: Cambridge University Press.

MacWhinney, B. and O'Grady, W. (eds.). 2015. *The Handbook of Language Emergence*. Oxford: Wiley-Blackwell.

McMahon, A. M. S. 1994. *Understanding Language Change*. Cambridge: Cambridge University Press.

Milroy, J. 2003. "On the role of speaker in language change." In Hickey 2003: 143–161.

Milroy, J. 2006. "Language change and the speaker: on the discourse of historical linguistics." In Cravens 2006: 147–165.

Milroy, J. and Leslie, M. 1985. "Linguistic Change, Social Network and Speaker Innovation." *Journal of Linguistics* 21 (2): 339–384.

Montefiore, A. C. R. G. and Noble, D. (eds.). 1989. *Goals, No Goals and Own Goals*. Unwin Hyman. Republished by Routledge, 2021.

Mufene, S. 2001. *The Ecology of Language Evolution*. Cambridge: Cambridge University Press.

Nedergaard Thomsen, O. (ed.). 2006. *Competing Models of Linguistic Change, Evolution and Beyond*. Amsterdam: John Benjamins.

Nerlich, B. 1990. *Change in Language: Whitney, Breal, and Wegener*. London and New York: Routledge.

Noble, D. 2006. *The Music of Life: Biology beyond Genes*. Oxford: Oxford University Press.

Noble, R., and Noble, D. 2018. "Harnessing stochasticity. How organisms make choices." *Chaos* 28, 106309. https://doi.org/10.1063/1.5039668.

Ohala, J. 1989. "Sound change is drawn from a pool of synchronic variation." In Breiviket et al. 1989: 173–198.

Ohala, J. 1992. "What's cognitive, what's not, in sound change." In Kellermann and Morrissey 1992: 309–355.

Poplack, S. and Torres Cacoullos, R. 2015. "Linguistic emergence on the ground: A Variationist paradigm." In MacWhinney and O'Grady 2015: 267–291.

Pütz, M (ed.). 1992. *Thirty Years of Linguistic Revolution. Studies in Honour of Rene Dirven on the Occasion of his 60th Birthday*. Amsterdam: John Benjamins.

Ritt, N. 2004. *Selfish Sounds and Linguistic Evolution: A Darwinian Approach to Language Change*. Cambridge: Cambridge University Press.

Rosenbach, A. 2008. "Language change as cultural evolution: Evolutionary approaches to language change." In Eckart et al. 2008: 23–75.

Sapir, E. 1921. *Language: An Introduction to the Study of Speech*. Harcourt, Brace and World. Ltd.

Saliloko, M. 2001. *The Ecology of Language Evolution*. Cambridge: Cambridge University Press.

Taylor, C. 1964. *The Explanation of Behaviour*. London: Routledge.

Tomasello, M. 2008. "Why don't apes point." In Eckardt 2008: 375–395.

Trudgill, P. 2008. "Colonial dialect contact in the history of European languages: On the irrelevance of identity to new-dialect formation." *Language in Society* 37 (2): 241–254.

Weinreich, U., Labov, W. and Herzog, M. 1968. *Empirical Foundations for a Theory of Language Change*. Austin: University of Texas Press.

Wilkes, K. 1989a. "Representation and Explanation." In Montefiore and Noble 1989: 159–185.

Wilkes, K. 1989b. "Explanation – How Not to Miss the Point." In Montefiore and Noble 1989: 194–211.

Zlatev, J. et al. (eds.). 2005. *The Shared Mind: Perspectives on Intersubjectivi*ty. Amsterdam: John Benjamins.

Croatian Journal of Philosophy
Vol. XXII, No. 66, 2022
https://doi.org/10.52685/cjp.22.66.5
Received: August 10, 2022
Accepted: September 22, 2022

Machine Learning, Functions and Goals

PATRICK BUTLIN
University of Oxford, Oxford, UK

Machine learning researchers distinguish between reinforcement learning and supervised learning and refer to reinforcement learning systems as "agents". This paper vindicates the claim that systems trained by reinforcement learning are agents while those trained by supervised learning are not. Systems of both kinds satisfy Dretske's criteria for agency, because they both learn to produce outputs selectively in response to inputs. However, reinforcement learning is sensitive to the instrumental value of outputs, giving rise to systems which exploit the effects of outputs on subsequent inputs to achieve good performance over episodes of interaction with their environments. Supervised learning systems, in contrast, merely learn to produce better outputs in response to individual inputs.

Keywords: Agency; machine learning; reinforcement learning; artificial intelligence; Dretske.

1. *Introduction*

One of the most powerful ideas in modern philosophy of mind is that an entity's origins can ground standards of success or evaluation to which its activities are subject. The relevant origins here are histories of learning or selection. This idea builds on the claim from philosophy of biology that selective history grounds biological function (Garson 2016) and has been prominently used in theories of representation (e.g. Millikan 1984, Papineau 1993, Shea 2018), as well as in teleofunctional theories of mental states (Sober 1985, Lycan 1987). In the theory of representation this idea helps to explain the correctness conditions which are deeply connected with meaning. In teleofunctionalism it helps to explain the fact that mental states and processes stand in normative

relations to one another—for instance, that it is part of the function of desires to cause motivation to act in combination with beliefs.

This idea may also be used in analysing agency. Agents engage in activity which is purposeful, functional, or otherwise governed by norms or standards, and their etiologies may ground these features. Glaciers interact with their environments but they are not agents because this activity is not governed by standards of correctness or evaluation. There is no sense in which glaciers aim to, or are supposed to, meet any such standards. Living organisms, in contrast, are at least candidates for agency, because much of their activity is purposeful or functional. I will say that agents engage in "norm-governed" activity, using the word "norm" very broadly to refer to non-arbitrary standards of correctness or of better or worse performance. Norm-governed activity is a necessary but not sufficient condition for agency, because the heart—for example—engages in activity which can be more or less successful according to its biological function, but the heart is not an agent. So agency is a species of which norm-governed activity is the genus.

Another way to see the point that agency is norm-governed is to start from the idea that agents pursue goals. If this is the case, agents' activity can be evaluated according to whether it helps to achieve their goals. Having a goal and having a function are two different ways to be subject to norms. In this paper, I will suggest that to have a goal, and thus to be an agent, it is necessary to have a history of learning or selection of a particular kind. Histories of this kind are made possible by certain capacities, and make others possible in turn. I will focus on formulating my claim in the context of a particular case; more work will remain to test the claim in other contexts.

My discussion will focus on the case of machine learning and in particular on the distinction between reinforcement learning and supervised learning. Machine learning researchers standardly refer to entities which undergo reinforcement learning as "agents", and reinforcement learning algorithms are designed to solve problems of the same general form of those which face biological agents (Sutton and Barto 2018). Furthermore, concepts and algorithms from reinforcement learning research are now widely used to explain value learning and action selection in humans and other animals (Niv 2009, Dolan and Dayan 2013). So it is natural and plausible to associate reinforcement learning with agency. In contrast, there are many systems trained by supervised learning, such as image classifiers, spam filters and translation tools, which do not seem to be agents. I will suggest an account of agency which vindicates these initial impressions, on the grounds that reinforcement learning is an example of the kind of process which gives rise to goals, but supervised learning is not.

This paper is therefore concerned with minimal agency—with the most basic distinction between those entities which are agents and

those which are not. It contrasts with much philosophical research on agency, which is concerned with the subtleties of human agency. Humans make plans, collaborate with others, experience emotions, and reflect on our own motives and choices, but none of these features seems to be essential to agency. I will start from a theory of minimal agency developed by Fred Dretske (1985, 1988, 1993, 1999), partly because it is abstract enough to be applied to the cases I am concerned with. I will set aside alternative approaches to minimal agency which are more specifically focused on the biological domain, such as those by Barandiaran et al. (2009) and Burge (2009).

In much of the paper I will not discuss the normative aspect of agency explicitly. After presenting Dretske's theory I will criticise it on the grounds that it implies that supervised learning-trained image classifiers are agents (section 2). I will then examine the differences between supervised learning and reinforcement learning, and propose a modification to Dretske's account, in section 3. In section 4 I will illustrate and elaborate on my proposal by discussing further examples of machine learning. In section 5, however, I will return to the idea that agency arises from histories of a particular kind, which give rise to entities which have goals and are consequently subject to associated norms. I will reformulate my proposal in these terms, building on the claim that natural selection gives rise to traits with biological functions.

2. *Dretske's theory of agency*

According to Dretske (1993, 1999), action is behaviour "controlled" or "governed" by thought. His account of agency forms part of his ambitious and elegant theory of intentionality and mental causation, which is presented in *Explaining Behavior* (1988) and several associated articles. One central claim of the account is that learning is necessary for agency. This learning must establish a structure in which a form of behaviour is produced selectively in response to features of the environment, through the operation of an internal state of the system. This internal state must be correlated with a feature of the environment, and must cause the behaviour partly in virtue of this correlation. That is, for some output of a system of type B to be an action, the following conditions must be met:

i. Internal states of the system of some type R are correlated with a feature of the environment E.
ii. The system learns to produce outputs of type B when in R-states.
iii. This learning happens in part because R-states are correlated with E.

For a system as a whole to be an agent, it must perform actions; a token output of type B is an action when it is caused by an internal state of type R through the route established by learning.

For example, consider a bird which learns to eat red pellets. For this to happen, the bird must have a visual system which enters a state of a certain kind when red pellets are in its field of view. If it pecks at and eats red pellets in the course of exploring its environment, and this behaviour is rewarded (e.g. because the pellets are palatable), it may learn to eat them selectively. This will involve a causal connection being formed between the visual system state that is correlated with red pellets and the behaviour of pecking and eating. This process will result in an arrangement which satisfies Dretske's conditions, and hence, according to Dretske, in the bird's becoming disposed to perform the action of eating red pellets.

In this case, the visual system state would not merely carry information about the presence of red pellets, but would come to be used as an indicator of red pellets. For Dretske, this means that it would come to represent the presence of red pellets. Alternatively, as he also puts it, it means that being in this internal state would amount to the bird's "thinking", or "believing", that red pellets are before it.

This "thought" or "belief" would then cause the behaviour of pecking and eating. For Dretske, a crucial point is that it would cause this behaviour in virtue of its content (behaviour being caused by thought is not enough, because this could happen without content being relevant). This would be the case because the correlation between the state and the presence of red pellets—the relation that underlies content—would have been a contributing cause of the connection's being established between the state and the behaviour. We would have a case of behaviour governed by thought, and therefore of agency.

As an influential account of content, mental causation and agency this picture has naturally been criticised.[1] One important criticism offered by Dennett (1991) is that it is not clear why the relationship between environmental conditions, internal states and behaviours must be established by learning rather than by evolution or design. A simple but unsatisfying response is that plants and simple artifacts would count as agents without the learning requirement. Thermostats are constructed so as to have internal states which correlate with low temperatures, which cause heating-activation outputs. The scarlet gilia, a plant which Dretske (1999) uses as an example, has flowers which change colour at the height of summer. It must therefore have some internal state which is correlated with the season, which is a proximal cause of this change. But in neither case is it appealing to say that the system is an agent, or that its output is "governed by thought". Some further justification might be achieved by saying that agents must be "autonomous" in the sense of Russell and Norvig (2010)—that is, that they must have a degree of independence from the knowledge of their designers, or more generally from the information which contributed to their initial forms. There is more to be said to fully justify the learning

[1] For criticisms which I will not discuss here, see Hofmann and Schulte (2014).

requirement, but here I will grant Dretske the point, in order to concentrate on a different feature of his theory.

I claim that Dretske's theory is insufficiently demanding because it entails that certain supervised learning-trained systems are agents.[2,3] For example, consider AlexNet (Krizhevsky et al. 2012), an image classifier using a deep convolutional neural network which was one of the defining advances of the development of deep learning. AlexNet is trained to label images as belonging to one of 1000 categories, in the following way. First an image is drawn from the training set and given to AlexNet as an input. This causes the network to produce some output, which takes the form of an assignment of probabilities to each of the categories. The correct label is provided, and the system uses gradient descent and backpropagation to adjust the network weights. This process is then repeated with further images from the training set, and the adjustments gradually increase the likelihood that the network will assign the highest probability to the correct label.

This may reasonably be described as a process of learning. The system undergoes endogenous, systematic changes in response to feedback which improve its performance, and it does so because it has been designed to change in this way. Furthermore, this learning seems to result in a situation which satisfies Dretske's criteria. After it has received some training, patterns of node activation in AlexNet will be correlated with type of input image—there may be some particular pattern correlated with images of pandas, for example. These patterns will cause AlexNet to produce particular kinds of outputs. The "panda" pattern will cause outputs which assign high probability to the "panda" category, and low probability to other categories. This situation will arise because the "panda" patterns are correlated with images of pandas, so weight combinations through which these patterns cause "panda" outputs will tend to be preserved. So AlexNet learns to produce outputs selectively in response to features of its environment, via internal states which indicate these features.

This is a problem for Dretske's account because AlexNet does not pursue any goal, and is not naturally described as an agent. It performs the function of classifying images, but not every entity which performs a function is an agent (as illustrated by the case of the heart). In the next section I will contrast supervised learning with reinforcement learning, which will allow me to give a more detailed analysis of this case.

[2] Strikingly, Dretske (1993) writes that genuine artificial intelligence is impossible, because being artificial is incompatible with being a product of learning, and the latter is necessary for genuine intelligence. This is surprising because he mentions learning in connectionist systems in *Explaining Behavior*.

[3] "Systems" here refers to particular implementations of algorithms—in this case, algorithms generated by the operation of implementations of further, supervised learning algorithms. Throughout this paper, when I suggest that artificial systems could be agents, my claim concerns implementations, not algorithms.

3. *Supervised learning, reinforcement learning and agency*

Machine learning problems and techniques are generally taken to belong to one of three classes: unsupervised learning, supervised learning and reinforcement learning. I consider only the latter two here, leaving unsupervised learning aside. In this section I describe supervised learning and reinforcement learning, then identify a difference which matters for agency.

According to Russell and Norvig's standard textbook on AI (2010: 695),

> The task of supervised learning is this:
>
> Given a training set of N example input-output pairs
>
> $(x_1, y_1), (x_2, y_2), \ldots (x_N, y_N),$
>
> Where each y_j was generated by an unknown function $y = f(x)$, discover a function h which approximates the true function f.

AlexNet is an example of a solution to a task of this form, because there is some function which takes each image in a labeled set to the correct label. An artificial neural network such as AlexNet can be seen, at each stage of training, as realising whatever function describes the transitions it is disposed to make from inputs to outputs. As AlexNet is trained this function comes to more closely approximate the true, target function.[4]

There are two noteworthy features of supervised learning which help to distinguish it from reinforcement learning. These both arise from the form of the training set, as a non-ordered set of input-output pairs. First, the feedback which the learning system receives, which drives its learning, specifies the correct output for the input just provided. Second, the input which is provided on each occasion and the correct output for that input are independent of any other actual or potential inputs or outputs. In particular, the system's outputs do not affect subsequent inputs.

Russell and Norvig define reinforcement learning as follows (2010: 830):

> The task of reinforcement learning is to use observed rewards to learn an optimal (or near optimal) policy for the environment.

Rewards are a form of feedback in which a numerical signal, which can have a positive, negative or zero value, is provided to the learning system after it produces each output. In reinforcement learning the next input (which is called a "state") depends probabilistically on the previous one and the system's output (called an "action"). The optimal policy for the environment is defined as that which maximises expected cumulative reward.

[4] For more on convolutional neural networks such as AlexNet, see Buckner (2019).

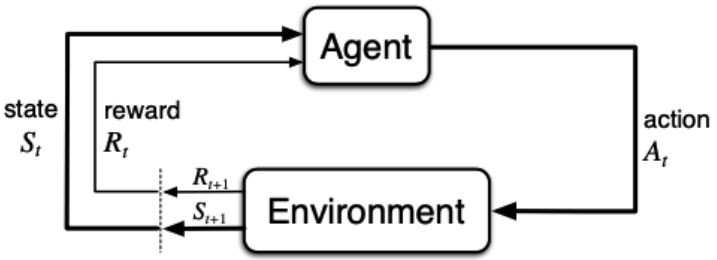

Figure 1. Illustration of reinforcement learning from Sutton and Barto (2018).

This arrangement is illustrated in figure 1. Here the "agent" is the system which undergoes reinforcement learning. At each time-step the system receives the state of the environment as input, produces an action as output, and receives a reward and an observation of the new state. In reinforcement learning environments are made up of transition functions, which describe the probabilities of new states given prior states and actions, and reward functions, which describe how much reward the agent will receive after each action.

An important advance in reinforcement learning from roughly the same period as AlexNet combined deep neural networks with a method called Q-learning to achieve human-level performance on Atari games (Mnih et al. 2015).[5] We can call this system DQN (for "Deep Q-Network"). As all reinforcement learning systems do, DQN receives both observations of the state of the environment and reward. Observations of the state of the environment take the form of maps of pixel values making up what would be displayed on a screen for human players, and reward is constituted by the game score. Outputs are actions possible for human players, such as producing the in-game effect of pressing a joystick button. DQN is trained separately on each game, losing its capacity to play one when trained on another.

To understand how DQN works, the most important element is the Q-learning algorithm. The function *Q(s, a)* is the action value function for the environment, describing how much cumulative reward can be expected to follow from taking action *a* in state *s* (which also depends on the system's policy, i.e. the actions it will subsequently choose). This function is somewhat analogous to the target function *f* in supervised learning, in that a reinforcement learning system will behave optimally if it always selects the action that maximises the Q-function for the current state. Analogously to AlexNet, DQN's outputs are determined by maximising its current estimate of the Q-function. There is the significant difference, though, that DQN is not given the true Q-value for the action it has just taken. Instead, it observes only the immediate change in the game score. This is very different, because—for example—an ac-

[5] The description given here is simplified in significant respects; see Mnih et al.'s paper for more details.

tion may cause no immediate change in the score, and yet be necessary to reach a state from which the highest scores are accessible.

Nonetheless, it is possible to use reward feedback to reach an approximation of the true Q-function. The method is to update estimated Q-values in the direction of the temporal difference error, given by the following formula:

$$R + \gamma\, Q(s', a') - Q(s, a)$$

Here R is the reward, γ is a discount factor, $Q(s', a')$ is the estimated value of the best action in the new state, and $Q(s, a)$—the value to be updated—is the estimated value of the action just taken in the previous state. The effect of this is that credit for rewards is passed back through the sequences of actions that lead to them.

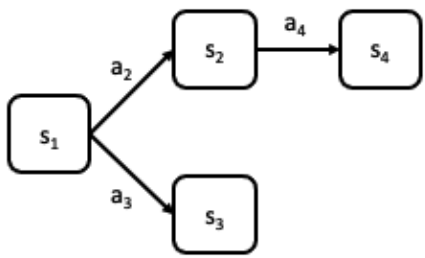

Figure 2. Illustration of Q-learning.

For example, consider the partial environment shown in figure 2, and suppose that the agent receives a high reward in s_4. In that case the temporal difference error for (s_2, a_4) is likely to be positive, so the agent's estimate of $Q(s_2, a_4)$ will be adjusted upwards. When the agent next performs a_2 in s_1, and thus reaches s_2, this higher value of $Q(s_2, a_4)$ will again likely mean a positive temporal difference error—because this will take the place of $Q(s', a')$ in the formula—so the agent's estimate of $Q(s_1, a_2)$ will be boosted. Credit for getting the high reward will be distributed back from a_4 to a_2 (and it could continue to be passed back in the same way). This could lead to the system forming a disposition to perform a_2 rather than a_3 in s_1 even if the latter led to greater immediate reward. In this way, the actions in sequences which lead to high rewards come to be represented as having high Q-values.

Reinforcement learning differs from supervised learning in each of the two features mentioned above. First, the feedback which drives reinforcement learning does not specify the correct output for the input just received. Instead, it is made up of an observation of the next state and a reward signal. Second, the identity of the next state is not independent of the previous one—instead, it is affected by the previous state and the action just performed. This means that reinforcement

learning systems engage in interaction with their environments—these are not just sources of inputs to which they must respond, but are also affected by their outputs in ways which affect their inputs in turn. In addition to these two features, in reinforcement learning there is a measure of success over episodes of interaction, and systems are equipped with algorithms which promote good performance on this measure. To perform well, in general, a reinforcement learning system must do more than just learn which actions yield most immediate reward. It must also learn how to reach states from which high levels of reward are available. That is, it must learn to exploit the fact that its outputs affect which inputs it will receive.

I propose that reinforcement learning systems are agents because, in addition to satisfying Dretske's conditions, they are capable of *instrumental* behaviour. To behave instrumentally is to produce outputs because these outputs contribute to good performance over episodes of interaction, such as by making it possible to access later rewards. Instrumental behaviour is both possible and necessary for reinforcement learning systems for the reasons just described. In particular, Q-learning and related methods produce instrumental behaviour because outputs come to be selected in virtue of their conduciveness to later rewards.

In contrast, AlexNet's outputs cannot be instrumental because they have no effect on subsequent inputs. Even if they did have an effect, the learning method employed in AlexNet is not sensitive to sequences of inputs, outputs and subsequent inputs, so it could not learn to engage in instrumental behaviour. The gradient descent algorithm by which AlexNet's weights are adjusted works by comparing the actual output for the current input with the correct output for that input. The feedback in supervised learning—that is, the information provided to the system which is affected by its outputs and which drives learning—does not include the identity of the next input. This difference between AlexNet and reinforcement learning systems makes sense because for AlexNet good performance overall just consists in producing the correct output for each input. For reinforcement learning systems, what makes outputs correct is how they contribute to maximising reward.

This view can be captured by the following claim about agency:

Instrumental view: An entity is an agent if and only if:

i. It produces some of its outputs selectively in response to inputs, as a result of a process which includes learning.

ii. This process is sensitive to instrumental value, where this means that it is influenced by information about input-output-input contingencies and functions to promote a specific form of feedback over episodes of interaction with the environment.

This view of agency combines two features: instrumentality in behaviour, and the learnt selectivity which Dretske describes. These two

features appear to be orthogonal, in that AlexNet learns to produce outputs selectively, but these are not instrumental, whereas a robot programmed to move efficiently through a specific maze would produce instrumental outputs without learning. However, it would be a mistake to think of my account as made up of separate instrumentality and learnt-selectivity conditions. Instead, what is crucial for agency is that the learning process is sensitive to instrumental value, so the system learns to produce outputs selectively because they contribute to good performance over an episode of interaction. One of the examples I will discuss in the next section will serve to illustrate this point.

4. *More on machine learning*

In this section I will discuss a series of further examples involving machine learning. Subsections 4.1 and 4.2 will cover other varieties of reinforcement learning, and add more detail to my account of how this form of learning is related to agency. Subsection 4.3 will discuss the possibility of using supervised learning to mimic optimal behaviour in a reinforcement learning-style environment; this case will prompt the clarification to my view suggested at the end of the last section. Finally, in subsection 4.4 I will comment briefly on agency in large language models.

4.1 *Varieties of reinforcement learning*

In the theory of reinforcement learning, a distinction is often made between "model-free" and "model-based" methods. The difference is that model-based methods involve the system learning and using a representation of the transition function, which can also be thought of as a model of the environment. Q-learning is a typical example of temporal difference learning, which is the most broadly-applicable form of model-free reinforcement learning. So in this subsection I will comment on varieties of reinforcement learning other than temporal difference learning, beginning by showing that systems which use typical model-based methods satisfy Dretske's conditions for agency and are capable of instrumental behaviour.

This claim can be illustrated by considering a model-based algorithm called R-Max (Brafman and Tennenholtz 2002). In this algorithm, look-up tables are maintained which store estimates of the transition function and reward function for the environment (the use of look-up tables means that this method is only suitable for finite environments). The rows in the transition function table record information about the new state which is expected following each action in each initial state, and the rows in the reward function table record the reward which is expected in each state. Actions are selected by exhaustive calculation of the cumulative reward of their expected consequences, looking ahead a fixed number of steps, with the action that begins the most rewarding sequence being chosen.

A system using this algorithm would satisfy Dretske's conditions because it would produce outputs selectively as a result of learning. After an initial period of exploration, such a system would develop dispositions to perform particular actions in particular states because its model would imply that these would lead to the greatest cumulative reward over the period to which its look-ahead extended. These actions would be caused by internal states correlated with states of the environment, and the causal links between internal states and actions would be explained by a combination of learning—which would establish the agent's model of the transition function and reward function—and reasoning—which would be used to select actions on the basis of the model.

Furthermore, the system would be capable of instrumental behaviour, because it would look ahead more than one step when selecting outputs. It would choose the actions which would allow it to maximise cumulative reward over multiple steps, meaning that its actions would be chosen for their contributions to good performance over episodes of interaction. The cases of temporal difference learning and model-based reinforcement learning illustrate that instrumental behaviour can be generated in different ways—either through learning algorithms which carry information about reward backwards through sequences of actions, or through action selection algorithms which use learnt models to look forward through such sequences.

A different form of model-free reinforcement learning is called Monte Carlo control (Sutton & Barto 2018). Monte Carlo control is notable because, whereas the sensitivity to instrumental relationships between actions and subsequent states is more explicit in R-Max than in Q-learning, this sensitivity is even less explicit in Monte Carlo control than in Q-learning. Monte Carlo control works in the following way. The system's purpose is to maximise reward in an environment with an end-state, which it engages with repeatedly (Monte Carlo control only works in cases like this). It starts by following some fixed policy many times, perhaps from a range of initial states. It records how much total reward it receives subsequent to each state-action pair on each occasion, then estimates Q-values for the policy it has been following by taking the mean of each set of observations. Then it improves its policy by choosing actions with higher Q-values, and repeats the process.

Monte Carlo control involves learning to select outputs for their contributions to cumulative reward, and hence involves exploiting the fact that outputs affect subsequent inputs. However, it does not depend on the agent's representing which states its actions lead to—either to feed into immediate updates as in Q-learning, or as part of the process of constructing a model. Instead, which states actions lead to influences how the system is updated by affecting the cumulative reward that follows actions. In this way, systems using this method are influenced by

information about instrumental relationships, so Monte Carlo control is sufficient for agency.

However, systems designed to solve two other problems studied in the context of reinforcement learning are not generally agents. These are the problem of planning, and multi-armed bandit problems (Sutton and Barto 2018). Planning is using a model of an environment which has been provided by the programmer to find an optimal policy. Planning is a crucial element of model-based reinforcement learning, but the capacity to plan does not suffice for agency, because it does not involve learning. Planners have little autonomy.

Multi-armed bandit problems are problems in which a number of outputs ("actions") are available to a system, each of which leads stochastically to a range of rewards, so that the system must learn which action is most rewarding. Systems for solving bandit problems are not generally agents, however, because the state of the environment does not change. So learning quickly about the relative values of outputs and maximising cumulative reward does not depend on exploiting the effects of outputs on subsequent inputs.

4.2 *Reinforcement learning systems pre- and post-training*

A further feature of reinforcement learning systems which calls for clarification of my account is that they change over time. Their abilities to navigate particular environments are gained only gradually, with this process often starting from an initial condition in which they select outputs randomly. In addition to this, engineers sometimes train systems with reinforcement learning only up to the point at which they reach a certain level of performance. After this the systems operate in the environment using a fixed policy or model, learnt during the training phase.

Different approaches to theorising about agency would give different verdicts on the status of reinforcement learning systems pre- and post-training. An approach which distinguished agents from non-agents according to whether they have the capacity to learn to behave in the relevant way would claim that pre-training systems are already agents, but systems which have been "frozen" after training are agents no longer. Combining this approach with my proposal that sensitivity to instrumental value matters would yield the view that agents are those entities with the capacity to learn to produce outputs selectively for their instrumental value. However, an alternative approach might claim that agents are those entities which perform actions, and actions are those outputs which are caused in the right way. Although the former approach has some attraction, I favour the latter. For an output to be an action it must be produced because the system has undergone a process which includes learning and is sensitive to instrumental value. This entails that reinforcement learning systems become agents gradually as they learn, because learning gradually comes to play a greater

role in explaining their outputs. It also entails that post-training systems which can no longer learn are still agents, because they still produce outputs as the result of a process of the right kind.

This approach has two advantages. First, as I will explain further in section 5, it makes it possible to analyse agency as a form of norm-governed activity, with the existence of the relevant norms grounded in history. Second, it is based on an analysis of actions as outputs which are caused in a certain way, and therefore subject to a certain form of explanation. It makes sense to use an account of action as the basis for a theory of agency, both because their capacity to perform actions is what is interesting about agents, and because not all outputs of agents are actions, so a substantive theory of action is needed in any case.

It may be objected at this point that I have not considered the possibility of an account of action which is based on proximal causes, such as reasoning which takes place "in the moment", rather than on the more distal role of learning. An account of this kind would avoid the potentially troubling implication of my view that a relatively long history is required.[6] One problem with accounts of this kind, however, is that they seem to have trouble distinguishing between AlexNet and DQN. Neither does much reasoning about which output to produce in response to a given input, but they still produce their outputs for very different reasons, and closer inspection of these shows important commonalities between DQN and model-based systems which do engage in in-the-moment reasoning.

4.3 *Mimicing agents using supervised learning*

It is sometimes argued that reinforcement learning is not necessary for agency on the grounds that it is possible to train a system by supervised learning that will mimic the behaviour of any reinforcement learning agent. The optimal policy for an environment is a function from states to actions, so if we know this function we can train a system to approximate it by supervised learning. More generally, if we know how a given reinforcement learning system will behave in a given environment, we can describe its behaviour as a function from states to actions, and again use supervised learning to train a system to mimic it. I claim that supervised learning systems of this kind are not agents, because—as I have just argued—the status of an entity as an agent depends on its history, not just on its current dispositions.

[6] A theory according to which a history of learning is required for agency faces the objection that a "swampman"—that is, a perfect replica of a living, adult human which emerges by chance from a swamp—would not immediately be an agent. I think this is the correct verdict on this case (see e.g. Millikan 1996, Shea 2018). See also McKenna (2016) and Zimmerman (2003) for discussions of other aspects of the role of history in agency.

This case is notable because it shows that learnt selectivity and instrumentality need to be combined in the right way to yield an attractive theory of agency. Dretske's theory entails that the status of an entity as an agent depends on its history because it requires an agent's dispositions to be a product of learning. However, we have already seen that Dretske's theory entails that supervised learning systems can be agents, so appealing to this theory alone will not justify a denial of agency in the present case. In addition to this, there is a sense in which the supervised learning "mimic" performs outputs for their instrumental value, because it is this value that explains why the reinforcement learning system performs them, or why they form part of the optimal policy. So neither Dretske's conditions nor instrumentality alone distinguishes the system trained by supervised learning from the reinforcement learning agent which it mimics.

What does distinguish these two systems is that in reinforcement learning, the learning and reasoning that combine to determine the system's policy are themselves sensitive to instrumental relationships. This sensitivity plays a role in the development (and thus, later, the causal history) of these systems, and thus contributes to explaining their actions. In the supervised learning case the learning process is insensitive to such relationships, which explain their actions only in so far as they play a role in the origin of the training data. One way to describe the difference is that in the supervised learning case talk of instrumental value would merely be an interpretative gloss on the meaning of the target function, while in the reinforcement learning case sensitivity to this value is built into the algorithm.

4.4 *Large language models*

I now turn to a final example, which is Transformer-based large language models such as GPT-3 (Brown et al. 2020) and PaLM (Chowdhery et al. 2022). The basic form of these systems is as "foundation models" for language (Bommasani et al. 2021), which are trained on large quantities of data to predict the next word from a given sequence. This can be described as "self-supervised" learning because the data does not need to be labeled by humans. However, it is very like the supervised learning discussed so far. The system trains itself by generating a prediction for the next word, then observing the actual next word and using the difference to calculate weight updates. So the feedback that drives learning specifies the correct output for the previous input. Furthermore, whether the system samples inputs at random from a corpus or works its way through systematically, in the course of training its outputs do not affect its inputs.

Foundation models trained in this way on enough data, using the Transformer network architecture, are capable of producing remarkably fluent language and performing challenging linguistic tasks (Brown et al. 2020, Chowdhery et al. 2022). Their capabilities are

sometimes further enhanced by various forms of fine-tuning, including by reinforcement learning. For example, InstructGPT (Ouyang et al. 2022) is based on GPT-3 but fine-tuned by reinforcement learning for generic good performance in responding to prompts, as judged by human users.

Foundation models are not agents because they do not learn to produce outputs for their instrumental value. In training their outputs do not affect their future inputs, so it is impossible for them to learn to exploit such effects. This point is obscured by the way in which foundation models are often used, which is to extend prompts by many more words, so as to generate texts of dozens or hundreds of words. When they are used in this way, foundation models' outputs are immediately added to their inputs, so this is a situation in which agent-like capabilities could be useful. But a language-producing system cannot produce individual outputs for the sake of facilitating subsequent outputs unless it has been subject to training in which its outputs affected subsequent inputs, and unless it has a way to evaluate sequences of outputs.

A complication to this picture is that some language models, such as those used for sentence-to-sentence translation, use an algorithm called "beam search" (Sutskever et al. 2014). One way in which a translation system might work would be to select words to output one by one, based only on their probabilities conditional on the input and on previous words. However, it is intuitive that such a system would be outperformed by one which internally generated a sample of complete sentences and compared their relative probabilities, before committing to any output. This is what beam search involves: starting from a small number of likely first words, the algorithm explores branches of the trees of possible sentences that begin with those words. Beam search is not sufficient for agency, however, because in the translation case the outputs of the system are whole sentences, and they are not selected for their effects on future inputs. It may be possible for foundation models to learn to do something like beam search in the course of selecting their outputs—to select words partly by looking at which words could follow them—but even this would not be agency if it was done solely as a means to maximising the likelihood of the next word, as opposed to influencing subsequent inputs.

Although they are not agents, foundation models are noteworthy because Transformers seem especially well-suited to learning to predict the next item in a sequence. This means that they can be used to learn to model environments and to predict the consequences of their actions. This is not sufficient for agency, but it is a crucial step along one route to agency—the model-based method for selecting actions for their instrumental value. For example, consider a hypothetical chatbot based on a foundation model trained on human dialogue. This chatbot might be good at predicting how a human user would respond to some output, and thus how that output would affect the state of the conversation, making new outputs and subsequent responses possible.

Its predictive capacity would enable it to take instrumental actions, provided that it could also evaluate possible future conversation states and combine these abilities in action selection.

5. *Selection, functions and goals*

So far in this paper I have focused on descriptive differences between reinforcement learners and supervised learners. I have proposed that only reinforcement learners perform actions, because only their outputs are the result of processes which are sensitive to instrumental value. However, agency can also be seen—as I suggested in the introduction—as a species of norm-governed activity (again, understanding norms merely as non-arbitrary standards of success or correctness). A potential advantage of my account of agency is that the differences in history which matter for agency could ground normative differences. This is the idea which I will develop in this section.

The idea that an entity's history can give rise to norms to which its activities are subject is exemplified by the selected-effects theory of biological function (Garson 2016). This theory, which is a mainstream view in the philosophy of biology, claims that if a component of some organism exists because it was selected for a certain activity, the function of the component is to perform that activity. This means that the activities of the component are subject to a norm; the component may either function correctly or malfunction (or perhaps it may function better or worse, according to a standard derived from its selective history). Building on this claim, and following other authors, I will argue that learning, as well as selection, can give rise to norms governing the activities of the entities which these processes modify. I will then propose that processes of learning or selection can give rise to different kinds of norms. As well as grounding the functions of components or traits, such processes can also give rise to goals, which entail norms governing the activities of whole systems.

The central idea of the selected-effects theory is that functions arise from "consequence etiology" (Shea 2018). In natural selection, traits with effects which contribute to greater reproductive success tend to persist and proliferate in populations, while those with other effects tend to die out. This means that natural selection is one context in which we can explain why traits exist by citing their effects—or, more precisely, the effects of prior tokens of their type—and therefore a context in which a form of teleological explanation is consistent with naturalism. Learning is like selection in this respect, because it involves the persistence of phenomena which have effects of the right kind. Training neural networks involves preserving those combinations of weights which have the right effects, and modifying those which have the wrong effects; and reinforcement learning involves only repeating those actions which contribute, through their effects, to greater cumulative reward.

However, not all situations in which something exists or persists as a result of its effects seem to give rise to functions. This was roughly the theory of function proposed by Wright (1973), and many apparent counterexamples have been proposed. For example, a leak in a gas hose may persist because the gas poisons anyone who tries to repair it (Boorse 1976).

Rather than attempting to defend a more restrictive general theory of functions, Shea (2018) argues for a disjunctive account. He claims that natural selection and learning from feedback are both ways in which a feature can come to persist (be "stabilised") in a population or system in virtue of its effects, which are such that the feature will then have the function of bringing about those effects.[7] His rationale for including learning is that, like natural selection, it is a means by which complex systems are developed and modulated in nature which make it possible for organisms to bring about outcomes robustly, especially by using representations. Shea's project is to justify appeals to representation in explanations in cognitive science, and he claims that this is justified by the frequency with which we observe a certain abstract pattern: apparently-representational features are stabilised by natural selection and learning in the service of producing outcomes robustly.

This paper is not concerned with representations and focuses on non-biological learning. However, it remains true that learning from feedback, like natural selection, is a form of consequence etiology which can give rise to complex and cumulative adaptations and which enables systems to produce outcomes robustly. Furthermore, learning is—in all real cases—itself a trait which has origins either in natural selection or in the design of artifacts by intelligent agents. Forms of learning themselves have functions. This should give us greater confidence in attributing norms in this context.

The analogy between natural selection and learning from feedback is not perfect, but to the extent that there is an analogy, these processes map onto one another in the following way. Natural selection acts on populations, while learning acts on "systems"—including human and animal minds and computer systems of various kinds. In natural selection, traits of organisms become more or less prevalent in populations, with some becoming near-universal for extended periods, while in learning features of systems such as behavioural dispositions or combinations of network weights are preserved or modified, with some stabilised. Stability in both cases is a consequence of stable features of the environment. Reproduction and persistence or modification are both determined by feedback. In natural selection, organisms bear combinations of traits, these traits have effects on the environment, and these effects determine how many offspring the organisms will produce, thus causing traits to become more or less prevalent in the population. In

[7] Shea also claims that contributions to the persistence of an organism can ground functions, but this is less relevant to the issue at hand.

learning, systems produce outputs, these prompt feedback from the environment, and this feedback determines which features of the system will persist or be modified. The state of the environment which faces a new generation in the case of natural selection is analogous to the input to a learning system, and the traits of that generation are analogous to the features that determine the system's output.

Norm-generating processes of selection and learning therefore have the following five elements: an entity with features which are preserved or modified (a population or system); an environment; inputs from the environment to the entity; outputs with effects on the environment (with input-output transitions being determined by the features); and feedback from the environment, which determines which features are preserved or modified. This account gives us an abstract framework within which functions and goals, and the processes which give rise to them, can be described.

Functions arise when features of the entity which is affected by selection or learning are stabilised. The function of a stabilised feature of an organism or system is to perform the activity, or bring about the effect, that caused it to be stabilised. The effects of features cause them to be stabilised when they contribute to bringing about the right kind of feedback.

In contrast, systems come to have goals only in much more specific circumstances. What is crucial is how feedback leads to persistence and modification. In reinforcement learning, feedback consists of both reward and the next input. The system stores information about relationships between inputs and subsequent feedback, and uses this information in determining how to modify its features. Furthermore, these modifications follow rules which, in most environments, make a particular kind of feedback (greater reward) more likely. When these elements are in place, it is not only possible to explain the existence of features of the system in terms of the effects of their type, but also to explain some of the system's outputs in terms of the contributions that they tend to make bringing about greater reward over episodes of interaction with the environment. This kind of explanation involves attributing goals to whole systems, because it is whole systems which interact with environments across episodes, by producing sequences of outputs. Systems with goals also have features with functions, but entities with functional features do not always have goals, because they are not all formed by processes which respond to feedback in this specific way.

This account of goals is intended to be equivalent to my account of agency; all and only agents have goals in this sense. The systems with goals are those that perform actions, because actions are outputs that have been selected for their contributions to greater cumulative reward over episodes of interaction.

To test my proposal it would make sense to examine how it applies to biological cases. If the proposal implied that most animals are agents

while most other organisms, populations and sub-organismic systems are not, this would be some evidence in its favour. If it had other implications this might be evidence against. However, for this purpose it would be important to bear in mind that the account of goals which I have just offered is not intended to describe what it is for a person to have a goal in mind when performing an action, or for an animal to behave in a goal-directed way (as opposed to habitually; Dolan & Dayan 2013). Talk of goals and goal-directedness is widespread and these terms are used in several ways. Instead, I have offered an account of goals which is intended to mark a distinction between the norms governing agency and those governing other forms of activity. This is just one of the ways in which human activities can have goals.

6. *Conclusion*

I have argued that to be an agent an entity must come to produce outputs for their instrumental value. For this to be the case, the agent's dispositions must arise from processes of learning or reasoning which are sensitive to instrumental value. That is, the modifications that arise in agents as a result of feedback from the environment must be modulated by information about relationships between outputs, inputs and subsequent reward. One source of support for this account comes from the idea that agents characteristically pursue goals. This means that an agent's individual actions must be subject to standards of success according to their conduciveness to the agent's goals. The existence of such norms could be explained by the operation of learning and reasoning processes of the kind just described.[8]

References

Barandiaran, X., E. Di Paolo and Rohde, M. 2009. "Defining agency: Individuality, asymmetry, normativity and spatio-temporality in action." *Adaptive Behavior* 17: 367–386.

Brafman, R. and Tennenholtz, M. 2002. "R-Max: A general polynomial time algorithm for near-optimal reinforcement learning." *Journal of Machine Learning Research* 3: 213–231.

Bommasani, R. et al. 2022. "On the opportunities and risks of foundation models." *arXiv* preprint.

Boorse, C. 1976. "Wright on functions." *Philosophical Review* 85: 70–86.

Brown, T. et al. 2020. "Language models are few-shot learners." *arXiv* preprint.

Buckner, C. "Deep learning: A philosophical introduction." *Philosophy Compass* 14 (10).

Burge, T. 2009. "Primitive agency and natural norms." *Philosophy and Phenomenological Research* 79: 251–278.

[8] Acknowledgements: I would like to thank Robert Long, Steve Petersen, Brad Saad, Derek Shiller and Jonathan Simon, and the participants at the Kathy Wilkes Memorial Conference, for their help with this paper.

Chowdhery, A. 2022. "PaLM: Scaling language modeling with Pathways." *arXiv* preprint.

Dennett, D. 1991. "Ways of establishing harmony". In McLaughlin (ed.). *Dretske and His Critics*. Oxford: Blackwell.

Dolan, R. and P. Dayan. 2013. "Goals and habits in the brain." *Neuron* 80 (2): 312–325.

Dretske, F. 1985. "Machines and the mental." *Proceedings and Addresses of the American Philosophical Association* 59: 23–33.

Dretske, F. 1988. *Explaining Behavior: Reasons in a World of Causes*. Cambridge: Bradford Books.

Dretske, F. 1993. "Can intelligence be artificial?" *Philosophical Studies* 71 (2): 201–216.

Dretske, F. 1999. "Machines, plants and animals: The origins of agency." *Erkenntnis* 51: 523–535.

Garson, J. 2016. *What Biological Functions Are and Why They Matter*. Cambridge: Cambridge University Press.

Hofmann, F. and Schulte, P. 2014. "The structuring causes of behavior: Has Dretske saved mental causation?" *Acta Analytica* 29: 267–284.

Krizhevsky, A., Sutskever, I. and Hinton, G. 2012. "ImageNet classification with deep convolutional neural networks." *Communications of the ACM* 60: 84–90.

Lycan, W. 1987. *Consciousness*. Cambridge: The MIT Press.

McKenna, M. 2016. "A modest historical theory of moral responsibility." *The Journal of Ethics* 20: 83–105.

Millikan, R. G. 1984. *Language, Thought and Other Biological Categories*. Cambridge: The MIT Press.

Millikan, R. G. 1996. "On swampkinds." *Mind and Language* 11 (1): 103–117.

Mnih, V. et al. 2015. "Human-level control through deep reinforcement learning." *Nature* 518 (7540): 529–533.

Niv, Y. 2009. "Reinforcement learning in the brain." *Journal of Mathematical Psychology* 53: 139–154.

Papineau, D. 1993. *Philosophical Naturalism*. Oxford: Blackwell.

Russell, S. and Norvig, P. 2010. *Artificial Intelligence: A Modern Approach* (3rd edition). London: Pearson.

Sober, E. 1985. "Panglossian functionalism and the philosophy of mind." *Synthese* 64: 165–193.

Sutton, R and Barto, A. 2018. *Reinforcement Learning: An Introduction* (2nd edition). Cambridge: The MIT Press.

Ouyang, L. et al. 2022. "Training language models to follow instructions with human feedback." *arXiv* preprint.

Wright, L. 1973. "Functions." *Philosophical Review* 82: 139–168.

Zimmerman, D. 2003. "That was then, this is now: Personal history v. psychological structure in compatibilist theories of autonomy." *Noûs* 37 (4): 638–671.

Croatian Journal of Philosophy
Vol. XXII, No. 66, 2022
https://doi.org/10.52685/cjp.22.66.6
Received: July 2, 2022
Accepted: October 12, 2022

Ascribing Proto-Intentions: Action Understanding as Minimal Mindreading

CHIARA BROZZO*
University of Barcelona, Barcelona, Spain

How do we understand other individuals' actions? Answers to this question cluster around two extremes: either by ascribing to the observed individual mental states such as intentions, or without ascribing any mental states. Thus, action understanding is either full-blown mindreading, or not mindreading. An intermediate option is lacking, but would be desirable for interpreting some experimental findings. I provide this intermediate option: actions may be understood by ascribing to the observed individual proto-intentions. Unlike intentions, proto-intentions are subject to context-bound normative constraints, therefore being more widely available across development. Action understanding, when it consists in proto-intention ascription, is a minimal form of mindreading.

Keywords: Action understanding; mindreading; Minimal Theory of Mind; intentions; normativity.

* I would like to thank Stephen Butterfill, Wayne Christensen, Paul Humphreys, Dunja Jutronić, Trenton Merricks, Bence Nanay, Krisztina Orbán, Chris Peacocke, Jonathan Schaffer, Joshua Shepherd, Barry C. Smith, Corrado Sinigaglia, Joulia Smortchkova, and especially Hong Yu Wong for feedback on previous versions of this paper. The research behind this article was supported by the FWO Odysseus grant G.0020.12N at the Centre for Philosophical Psychology (Universiteit Antwerpen), the Max Planck Society for the Independent Minerva Research Group, Space and Body Perception, led by Betty Mohler, the John Templeton Foundation (ACT Fellowship awarded to Hong Yu Wong), the Fritz Thyssen Foundation, and the Starting Grant ReConAg 757698 from the European Research Council awarded to Joshua Shepherd under the Horizon 2020 Programme for Research and Innovation.

1. *Introduction*

I watch you move your hand towards a teacup, and I understand that your movements are directed towards picking up that teacup. This is an instance of *action understanding*, namely the process by means of which someone identifies the outcome to which a series of movements are directed.[1] An *outcome* is here to be understood as a possible or actual state of affairs—for example, a teacup being picked up—that is the result of a series of movements. *Action* is here used interchangeably with *event*: there is no presupposition that, when the action is understood, it is understood as such—namely, as Anscombe (1957) and Davidson (1963) would have put it, as intentional under a description (see also Smortchkova 2018).

How do we understand other individuals' actions? Answers to this question tend to cluster around two extremes. On the one hand, it may be thought that actions are understood by ascribing to the observed individual a mental state representing the outcome being brought about (Goldman 2006 considers this possibility—see section 3). In the previous example, I would understand your movements as directed to the outcome of the teacup being picked up by ascribing to you, e.g., an *intention* to pick up the teacup, or to drink tea.[2] In other words, action understanding would be a form of *mindreading*, which is standardly conceived as the ascription of mental states—propositional attitudes, but also emotions—to others or to oneself (see, e.g., Stich and Nichols 1992; Goldman 2009).[3]

Connecting action understanding to standardly conceived mindreading requires that an observer engaged in action understanding is equipped with relevant mental state notions, such as that of intention. These notions may in principle be rather cognitively demanding, entailing, e.g., relations to many other mental states. For example, I may ascribe to you the intention to drink tea in conjunction with the intention to be a bit more awake, or to be a good host and keep me company in drinking tea, but not in conjunction with ascribing to you

[1] The notion of *directedness* is used to distinguish outcomes that are purposely brought about from those that are accidentally brought about. In the example just given, your movements are *directed* towards the outcome of the teacup being picked up. By contrast, if you moved in such a way as to *accidentally* spill the tea contained in the cup, it would not be the case that your movements were directed towards the outcome of the tea being spilled (see Sinigaglia and Butterfill 2015).

[2] The role of intention in the explanation of purposive behaviour has been amply discussed by Kathy Wilkes (see, for example, Wilkes 1989).

[3] Notice that this possibility about how action understanding works does not trivially follow from the definition of action understanding. This is because both the notion of outcome and that of directedness to an outcome are devoid of reference to mental states: a series of movements may be directed to a given outcome without there being any mental states representing that outcome. For example, the movements of a mechanical arm may be directed to the outcome of a teacup being picked up, without there being any mental state representing that outcome.

an abhorrence of tea, absent a further ascription of pressing reasons to nevertheless drink it.

On the other hand, at the other extreme, it may be thought that actions are understood in a way that does not draw on mental state ascription at all. Rather, action understanding is exhausted by relating observed movements to anticipated—and eventually observed—outcomes (Gergely and Csibra 1997, 2003; Roessler and Perner 2010; Spaulding 2013). This would make action understanding more similar to processes whereby most human adults understand physical interactions such as causal ones: to understand that a billiard ball has set another into motion through collision, no mental states are ascribed to either ball.

Lots of experimental research on action understanding, examples of which I shall illustrate later, suffers from the lack of an intermediate theoretical option. According to the two aforementioned options, actions are either understood through the ascription of cognitively demanding mental states, or without the ascription of any mental states at all. As sections 5.1 and 5.2 will show, neither option seems adequate in some cases of action understanding.

In answer to this impasse, this paper will put forward a proposal about what action understanding could involve that lies midway between the aforementioned two extremes (see also Andrews 2020). According to it, differently from the second extreme option, action understanding would involve the ascription of some mental states. Differently from the first extreme option, however, the mental states ascribed in understanding others' actions would not be as cognitively demanding as in full-blown mindreading.[4]

In the following sections, I shall, first, describe in detail the possibility that action understanding could consist in full-blown mindreading (sections 2–3). After that, I shall illustrate how action understanding might involve no mental state ascription (section 4). Making clearer the commitments of these extreme options will lay the ground for putting forward my own middle ground proposal (sections 5–ff.).

[4] The proposal I am going to put forward has analogous motivations as that made by Butterfill and Apperly (2013; see also Apperly and Butterfill 2009). According to the latter proposal, some creatures could ascribe to others mental states called *registrations*. These are like beliefs in some respects, but also simpler than beliefs, in a way that will be clarified in section 5. Butterfill and Apperly's proposal is motivated by the need to interpret certain findings in developmental psychology that the notion of belief is inadequate to explain (e.g., Onishi and Baillargeon 2005). The focus of my proposal, unlike Butterfill and Apperly's, is mental states that are like intentions in some respects, but also simpler than intentions. My proposal is independent of Butterfill and Apperly's: for reasons that will become clear later, the tenability of one does not hinge on the tenability of the other, and vice versa.

Also, my proposal assumes that mental states, whether minimal or full-blown, are representations of sorts, and will therefore not engage with anti-representationalist views of the mind (e.g., Gallagher and Hutto 2008). Lastly, my proposal is not to be seen as an alternative to either Simulation Theory (Gordon 1986; Heal 1986; Goldman 2006) or Theory Theory (Gopnik and Wellman 1992; Gopnik and Meltzoff 1997), as it is, in principle, compatible with both.

2. *Under what conditions would action understanding be mindreading?*

In order to present the first extreme option, according to which action understanding is full-blown mindreading, I want to clarify under what conditions action understanding would be mindreading.

Action understanding would *not* be mindreading in cases such as the following. Suppose that I observe Alice moving towards a teacup. Suppose that I understand Alice's movements as directed to the outcome of the teacup being picked up. Action understanding is complete at this point: the teacup being picked up has been identified as the outcome to which Alice's movements are directed. Then, subsequently, I additionally ascribe to Alice the intention to have a leisurely cup of tea. But this would not make the previous instance of action understanding an act of mindreading: the act of mindreading (ascribing to Alice the intention to have a leisurely cup of tea) would be *distinct* from that of action understanding (identifying Alice's movements as directed to the outcome of a teacup being picked up). In this example, mindreading begins when action understanding is already over. The moral of this example is that ascribing a mental state once an outcome has already been identified as that to which an observed series of movements are directed does *not* make action understanding an instance of mindreading. By contrast, action understanding would be mindreading if it involved ascribing mental states—either because mental state ascription is part of the process of action understanding, or because it is identical to it. For example, action understanding would be mindreading if ascribing to Alice the intention to have a leisurely cup of tea had a causal role in concluding that her movements are directed towards the outcome of the teacup being picked up.

3. *Intention ascription: full-blown mindreading*

Let me now present the option according to which action understanding would be full-blown mindreading, in line with the provisos offered in the previous section. I shall present one way for action understanding to be mindreading, which consists in ascribing an intention to an individual performing the action.

How does intention ascription relate to understanding actions? By virtue of the widely shared view that intentions represent or otherwise specify outcomes (see, e.g., Searle 1983; Bratman 1987). For instance, the intention to build a house represents the outcome of a house being built. Therefore, by ascribing an intention to build a house to an individual, one thereby identifies an outcome—the outcome represented by that intention—to which this individual's action is directed. It is a further question whether the observed individual actually *has* an intention to bring about the outcome (see Borg 2007; Sinigaglia 2008).

Of course, mental states other than intentions also represent outcomes—for example, beliefs and desires. Here I am assuming that the mental states ascribed in some cases of action understanding are intentions. Why? Because not only do intentions represent outcomes, and to this extent they are akin to beliefs, but, unlike beliefs, intentions represent outcomes with a world-to-mind direction of fit and a mind-to-world direction of causation (see, e.g., Searle 1983): in order for intentions to be fulfilled, the world has to conform to them, and intentions contribute to the required changes in the world. So, they are fit to fulfil the role of causes of the observed behaviour, unlike beliefs.

One could object that the same considerations about direction of fit and direction of causation make desires just as plausible candidate mental states to be ascribed in action understanding. Intentions and desires, however, differ in the following way. According to a standard conception, intentions are tools for planning. This is reflected in their being subject to characteristic normative constraints concerning consistency and rationality—in particular, what Bratman (1987) termed the *strong consistency requirement*. An intention satisfies the strong consistency requirement if and only if it is consistent with the rest of the subject's intentions, as well as with the rest of the subject's beliefs. It is a normative constraint in the sense that intentions should satisfy it in order to fulfil their role as tools for planning, but it is conceivable that intentions may break it (for example, I may intend to get ready in fifteen minutes all the while believing that it will take at least half an hour). If they do, then the subject is guilty of irrationality. No such normative requirement applies to desires (Bratman 1987; Holton 2009). In particular, having conflicting desires does not make a subject irrational. Due to the applicability of these normative constraints, intentions are better suited than desires to account for consistency relationships between ends and means that are recognised by certain subjects, as will be illustrated in the next section with the experiment by Gergely and colleagues (1995). For this reason, in what follows I shall focus on intentions as candidate mental states to be ascribed in the context of action understanding. I do concede that, if action understanding consisted in desire ascription, it would also be full-blown mindreading, but I leave a discussion of the case of action understanding consisting in desire ascription for another occasion.

Now I shall provide an example of how action understanding could be intention ascription, and therefore full-blown mindreading. This will consist in a specific version of the so-called *generate-and-test* model, introduced by Goldman as follows:

> The attributor begins with a known effect of a sought-after state, often an observable piece of behavior. He generates one or more hypotheses about the prior mental state or combination of states that might be responsible for this effect. He then "tests" […] these hypotheses by pretending to be in these states, feeding them into an appropriate psychological mechanism, and seeing whether the output matches the observed evidence. When a match is

> found […], he attributes the hypothesized state or combination of states to the target. (Goldman 2006: 45)

Action understanding involving intention ascription would take place if the generate-and-test model were instantiated with the following auxiliary assumptions. First, an observer hypothesises that the prior mental state responsible for the observed behaviour of another individual is an *intention* to bring about a certain outcome, and the observable behaviour consists in the bodily movements bringing about that outcome. Furthermore, the observer eventually does find a match between, on the one hand, the bodily movements that the hypothesised intention would produce and, on the other hand, the observed bodily movements. Therefore, the observer ascribes the hypothesised intention to the observed individual. Since the intention represents the outcome to which the observed movements are directed, this would be a case of action understanding. Therefore, one may understand an action by ascribing an intention.[5] This would make action understanding full-blown mindreading.

4. *Mere outcome identification: not mindreading*

A second extreme option concerning how action understanding could take place is without any ascription of mental states. I shall call this option *mere outcome identification*. Here is an example of it.

According to Gergely and Csibra (1997; 2003; Csibra and Gergely 1998), outcomes (which they call *goal-states*) are identified by one-year-olds thanks to the *teleological stance*. The teleological stance is an interpretational schema featuring three elements: an outcome, an action (a term that Gergely and Csibra use as synonymous with *a series of movements*—in line with its meaning in *action understanding*) and a set of situational constraints. An individual understanding an action by means of the teleological stance identifies all three elements, and moreover identifies actions as directed to (which may be read as *supposed to bring about*) certain outcomes. Between different actions directed to the same outcome, this individual is further capable of identifying the most rational action for bringing about the outcome given the current situational constraints (e.g., in the absence of an obstacle,

[5] The generate-and-test model is put forward within the framework of the Simulation Theory of mindreading, according to which an observer ascribes mental states to an observed individual by means of an attempt to replicate the workings of the latter's mind (Gordon 1986; Heal 1986; Goldman 2006; Gallese and Goldman 1998). This is reflected in the fact that, according to the generate-and-test model, the observer tests the hypothesised intention by *pretending to have that intention herself*. Notice, however, that intention ascription can take place outside of the simulationist framework. The generate-and-test model itself could be modified so as not include a commitment to the Simulation Theory, for example as follows. A subject hypothesises that the observed individual has a certain intention, and then draws on a theory about how intentions connect with ensuing bodily movements in order to make the relevant predictions about the movements that she should observe, were the observed individual to have the hypothesised intention.

approaching a target in a straight line is more rational than approaching it via a curved path). All this, according to Gergely and Csibra, is done without ascribing any mental states.

Here is an example of the teleological stance at work. In a violation-of-expectation study (Gergely et al. 1995), infants were habituated to a computer animation showing a small circle approaching a large one. In this animation, the small circle moved along a trajectory that looked like a jump, through which it approached the large circle while avoiding a rectangular obstacle. In the context of this computer animation, the outcome was the large circle being reached, the action consisted in the movements of the small circle, and the situational constraints consisted in the presence of the rectangular obstacle. After the infants had been habituated to this animation, they were shown two test displays. In both of them, the obstacle was removed. In one of the two test displays, the small circle approached the large circle in a straight line. In the other test display, the small circle approached the large circle following the same trajectory as in the habituation display, i.e. a trajectory that looked like a jump. Infants looked longer (which is taken to indicate surprise) at the latter test display than at the former.

Gergely and Csibra's (1997, 2003) interpretation is as follows. First, infants showed sensitivity to the directedness of movements to an outcome. In particular, they recognised that the small circle's movements were directed to the outcome of the large circle being reached. Furthermore, these infants recognised situational constraints—the presence of the obstacle in the habituation display or of an unblocked path in the test displays. Lastly, infants were capable of recognising the straight-line approach as more rational than the jump-like approach for bringing about the outcome (the large circle being reached) under the given situational constraints (an unblocked path). But this, according to Gergely and Csibra, happened without any representation (and, *a fortiori*, ascription) of mental states on the infants' part, and indeed there was no presupposition that the moving circle observed by the infant had a mind at all. On the contrary, the infant identifying a certain outcome being brought about is described as a *"mindblind" creature* (Gergely and Csibra 2003: 290).

It should at this point be clear that, if action understanding consisted in mere outcome identification (for example, in the form of the teleological stance; see also Perner and Roessler 2010), then action understanding would not be mindreading.

Let me take stock so far: I have singled out two extreme options for how action understanding could occur:

1. *Intention ascription*. An observer identifies an outcome to which a series of movements are directed by ascribing to an observed individual an intention representing that outcome. If action understanding consisted in intention ascription, it would be full-blown mindreading.

2. *Mere outcome identification*. An observer merely identifies an outcome to which a series of movements are directed, without ascribing any mental states. If action understanding consisted in mere outcome identification, it would not be mindreading.

Now I am ready to present my own proposal about how action understanding could occur, situating it midway between full-blown mindreading and the absence of mindreading.

5. *A third option: minimal mindreading*

Up to now, my discussion has been confined to two rather extreme options: in action understanding, either one ascribes to an observed individual an intention (so that action understanding is full-blown mindreading), or one identifies an outcome to which an action is directed (so that action understanding is not mindreading). I would now like to point out that the following middle ground should be explored: action understanding could be a minimal form of mindreading. A minimal form of mindreading occurs when minimal forms of mental states—i.e., mental states less cognitively demanding than propositional attitudes—are ascribed (see, e.g., Tomasello et al. 2003; Nanay 2013; Whiten 2013; Butterfill and Apperly 2013).

An example of a mental state less cognitively demanding than a propositional attitude is provided by *registrations*, postulated by Butterfill and Apperly (2013) as part of their proposed Minimal Theory of Mind. A registration is a relation between a subject, an object and a location. Like beliefs, registrations have correctness conditions insofar as an individual correctly registers an object at a location if and only if that object is actually at that location. Due to being relations rather than representations, however, unlike beliefs registrations are not sensitive to different modes of presentation of one and the same object (see Butterfill and Apperly 2013; Low and Watts 2013). Because of this, they are less cognitively demanding (and therefore more widely available within and across species) than beliefs. I will now provide some motivation for exploring minimal forms of intention.

5.1 *Some motivation for the ascription of minimal forms of mental states*

I will now illustrate two experimental results that I shall term unwilling vs. unable (Behne et al. 2005) and failed attempts (Meltzoff 1995). I will then explain how a possible interpretation of these results motivates considering the idea that action understanding could consist in the ascription of minimal forms of mental states—specifically, minimal forms of standardly conceived intentions.

Behne and colleagues (2005) tested infants (from 6 to 18 months of age) as follows. An infant faced an adult experimenter in the position to pass them an object. The infant was presented with both the following

kinds of scenario at different times: in one of these, the experimenter did not pass the object to the infant because the experimenter was *unwilling* to do so; in another, the experimenter did not pass the object to the infant because the experimenter was *unable* to do so (for example, the object slipped out of their hands). While 6-month-olds were not sensitive to the difference between unwilling vs. unable, infants from 9 months of age onwards were more *impatient* in the scenarios in which the experimenter was unwilling to pass them the object than in those in which the experimenter was well-meaning but clumsy, and therefore unable to pass them the object (compare Call et al. 2004 for a similar paradigm with chimpanzees).

There are at least two possible ways of accounting for the different reactions observed in the subjects of the above reported experiment: in terms of mere outcome identification or in terms of mental state ascription. According to an interpretation in terms of mere outcome identification, the infants from 9 months of age onwards identified the outcome to which the experimenter's action was directed. In particular, they understood the experimenter's movements in the *unwilling* condition as directed to the outcome of the object being withheld, and in the *unable* condition as directed to the outcome of the object being passed to them (though the experimenter failed to bring it about). According to an interpretation in terms of mental state ascription, the infants from 9 months of age onwards ascribed a *mental state* to the experimenter, one that represents the outcome to which the experimenter's movements are directed (object being withheld vs. object being passed to the infant). Absent any independent considerations, prima facie there is no reason to exclude an interpretation in terms of mental state ascription (see Michael and Christensen 2016 for doubts that interpretations of similar results in terms of mere outcome identification are adequate).

Now let me turn to failed attempts. In an experiment by Meltzoff (1995), 18-month-olds were shown failed attempts to perform a certain action (e.g., pulling apart a dumbbell-shaped toy) by an adult experimenter. When their turn came, these infants enacted the observed action, bringing about the outcome to which they interpreted it as being directed. They did not enact the observed action, however, when they were shown an inanimate object (a device with mechanical arms) executing the same movements as those performed by the experimenter.

As with unwilling vs. unable, two interpretations are possible. According to an interpretation in terms of mere outcome identification, 18-month-olds understood that the experimenter's movements and those of the inanimate object were directed to an outcome which failed to be brought about. With this first interpretation, the question arises as to why infants enacted the observed action bringing it to completion only in the former case, and not in the latter. According to an interpretation in terms of mental state ascription, 18-month-olds ascribed a mental state to the experimenter, one that represents the outcome to which the experimenter's movements are directed (e.g., pulling apart

the toy), but merely identified the outcome to which the observed movements were directed when faced with an inanimate object. It seems like an interpretation in terms of mental state ascription would have more explanatory power.

Once this option is on the table, the question now is: if infants do ascribe a mental state to the experimenter, what is the mental state in question?

5.2 *Why the ascription of standardly conceived intentions will not do*

In section 3, I explored the possibility that the mental state ascribed in action understanding could be a standardly conceived intention. But this does not look like a viable option for the experiment reported in the previous section. Why not? In section 3, I described intentions as states that are subject to normative constraints concerning consistency and rationality (Bratman 1987; Holton 2009). This makes intentions cognitively demanding: representing an intention implies being sensitive to the fact that this intention should not conflict with many of one's intentions and beliefs. Now, it is plausible to assume that the complexity of a mental state imposes constraints on the ease of identification and ascription of such mental states (see, e.g., Butterfill and Apperly 2013). Working on this assumption, and on the assumption that intentions are relatively complex mental states due to the normative constraints applying to them, it is highly implausible that creatures such as infants ranging from 9 to 18 months of age should be able to represent and ascribe standardly conceived intentions, insofar as this would place too high demands on their inferential abilities. Call this line of reasoning *can't have*.

Can't have makes it worthwhile to explore mental states representing outcomes to which actions are directed that are different from standardly conceived intentions—different insofar as their representation does not impose as high demands on infants' inferential abilities as standardly conceived intentions. In other words, these mental states should be such that infants between 9 and 18 months of age can represent and ascribe them.

Another consideration in favour of exploring this option, which I shall call *needn't have*, is that even creatures such as human adults, who could plausibly represent and ascribe standardly conceived intentions, *do not need* to ascribe anything as complex as that when they have to, e.g., tell someone who is unwilling to perform a certain bodily action apart from someone who is unable to do so.

In short, working on the assumption that we cannot rule out an interpretation of the above reported experiments in terms of mental state ascription, the *can't have* line of reasoning provides motivation for exploring the option that minimal forms of mental states could be ascribed in action understanding. Independent motivation for exploring this option is given by the *needn't have* line of reasoning.

6. *Proto-intentions: a minimal form of intention*

In this section, I will present a minimal form of intention that I shall call *proto-intention*. Just like intentions, proto-intentions are mental states with a world-to-mind direction of fit and mind-to-world direction of causation. However, differently from intentions, they represent outcomes in a less cognitively demanding way, modelled on a kind of outcome identification that, following Tomasello and colleagues (2005), I shall call *tracking the choice of plans*. In section 6.1, I will say what tracking the choice of plans is. In section 6.2, I will show how ascribing proto-intentions enables tracking the choice of plans while posing inferior cognitive demands on a subject's inferential abilities compared to ascribing standardly conceived intentions.

6.1 *Tracking the choice of plans*

Tracking the choice of plans consists in identifying an outcome to which an action is directed while also telling it apart from the specific means with which it was achieved. Several experiments can be taken to indicate that their subjects have the ability to track the choice of plans. One of them is the Gergely and colleagues' (1995) experiment described in section 4, where subjects can be interpreted as able to tell apart the outcome to which the observed movements are directed (i.e., the large circle being reached) from the means with which this is achieved (straight-line path vs. jump-like path).

Suppose that Gergely and colleagues' (1995) experiment, described in section 4, is interpreted as one in which action understanding involves some form of mindreading. I am leaving it open whether this is actually the case, but note that Gergely and colleagues themselves previously supported an interpretation of their own results in terms of mental state ascription (Gergely et al. 1995). Assuming an interpretation in terms of mental state ascription, what could the mental states ascribed by the infants be? My proposal is that they could be *proto-intentions*.

6.2 *Proto-intention ascription enables tracking the choice of plans*

In this section, I shall characterise proto-intentions as mental states partly analogous to intentions but subject to more local normative constraints concerning consistency and rationality.[6]

The way I shall characterise proto-intentions assumes that proto-intentions could be both *states one has*, i.e. that are part of someone's psychology, as well as *states one ascribes* to other individuals. This is one of the main differences between my proposed minimal forms of mental states and Butterfill and Apperly's registrations, described in

[6] I will focus on differences in normative constraints between proto-intentions and intentions, while leaving it open that they may differ also in other respects.

section 5: registrations are *not* supposed to be part of anyone's psychology, but rather useful tools for explanation and prediction on the part of the individual that ascribes them. In other words, registrations are supposed to be *states one ascribes*, but not necessarily *states one has*.[7]

Why think that proto-intentions could exist? Based on the idea that proto-intentions are both states one has and states one ascribes, support for the idea that some creatures might have proto-intentions comes from reflections on animal cognition made by Susan Hurley (2003).

Recall from section 3 that intentions are subject to characteristic normative constraints, such that one's intentions should (ideally) be consistent with the rest of one's beliefs and intentions. Notice that there is no principled boundary on the number of intentions and beliefs that one's intentions should not conflict with. Suppose, for example, that I intend to spend tomorrow writing a book chapter. Suppose that someone invites me to join them on a leisurely day out, walking in the countryside. Should I settle on that course of action, thereby forming an intention to spend tomorrow walking the countryside? A quick inference leads me to conclude that spending tomorrow walking the countryside means I will not do any writing. This conflicts with my intention to spend tomorrow writing a book chapter—which speaks in favour of not forming an intention to spend tomorrow walking the countryside. However, I also believe that it would be good for me to do some exercise—something that the doctor recently advised me to do, and I intend to follow his advice. Another inference leads me to conclude that spending tomorrow walking in the countryside would be the perfect way to follow my doctor's advice. But then I also believe that my book is overdue, which speaks in favour of my original intention… and so on. This is an illustration of how standardly conceived intentions presuppose in principle unbounded inferential abilities.

By contrast, Hurley (2003) pointed out that there is an interesting normative middle ground between the full-blown rationality that norms on intentions seem to require and, on the other hand, the complete absence of norms of rationality. In particular, according to Hurley, "[n]on-human animals can occupy islands of practical rationality: they can have *context-bound* reasons for action" (2003: 231, my emphasis). To make things more concrete, Hurley considers the following possibility, based on observations by Cheney and Seyfarth (1992), Tomasello (1999) and Tomasello and Call (1997). The possibility is that some animals (e.g., chimpanzees) could make transitive inferences in some

[7] A theory about how action understanding takes place that draws on states one has is amenable to being framed in terms of the Simulation Theory, in which the states one ascribes are precisely the states one has (I am here just pointing out the possibility of doing so, but I shall not pursue it in this paper). Note that, assuming that the states one ascribes are the states one has, there are deep and difficult questions about what enables subjects that have certain mental states to also ascribe them to other individuals (see, e.g., Tomasello 1999; Tomasello and Call 1997; Hurley 2003; Peacocke 2014).

contexts (e.g., social contexts) but not others (e.g., non-social contexts). For example, Hurley conjectures that a chimpanzee could make transitive inferences of the kind "A is dominant over B, B is dominant over C, therefore A is dominant over C" (where A, B and C are conspecifics), but not of the kind "A has more fruit than B, B has more fruit than C, therefore A has more fruit than C" (where A, B and C are trees). This would enable the chimpanzee to use the former, but not the latter, kind of information to guide their actions flexibly in relation to various goals. In other words, some animals' reasons for acting may be context-bound, that is, not generalise to all possible contexts. Inference to the best explanation would then make it plausible that, if actions whose reasons are context-bound could be driven by mental states, then these would have to be mental states that are subject to more local normative constraints than intentions. These are what I shall call *proto-intentions*: states with a world-to-mind direction of fit and mind-to-world direction of causation that are subject to a limited form of the strong consistency requirement. A limited form of this normative requirement merely prescribes that an individual's proto-intention should not conflict with:

(i) another proto-intention of that individual that is linked to the former via means-end reasoning, and
(ii) with information that the individual has about how one's end could be achieved in the circumstances (i.e., information about what Gergely and Csibra called *situational constraints*).

As an example, a proto-intention to reach another individual by following a straight-line path should be consistent both with one's intention to reach the other individual and with the information one has about the obstacles present on that path at the moment, and not with information spread across longer timescales (the latter illustrated in the previous example, concerning the intention to spend tomorrow walking the countryside).

At this point, I have introduced the notion of proto-intention to explain some cases of action production. Working on the assumption that proto-intentions are states that one has but also that one can ascribe, I shall now present the following possible way in which action understanding might occur:

3. *Proto-intention ascription.* One could identify an outcome to which an action is directed by ascribing to an observed individual a proto-intention representing that outcome. Given that proto-intentions are minimal forms of intentions, if action understanding consisted in proto-intention ascription, it would be a minimal form of mind-reading.

The reason why ascribing a proto-intention would be a useful strategy for identifying outcomes is that proto-intentions, just like intentions, represent outcomes with a world-to-mind direction of fit and have a mind-to-world direction of causation. They differ in the normative con-

straints to which they are subject. Intentions are subject to the strong consistency requirement, and, as a result, representing and ascribing intentions presupposes in principle unbounded inferential abilities—or at least, rather cognitively demanding inferential abilities. By contrast, proto-intentions are subject to a limited form of the strong consistency requirement, and, therefore, in order to represent and ascribe proto-intentions, one need only have inferential abilities that enable the evaluation of different potential means for achieving the same outcome. This makes proto-intentions in principle more widely available across development and species.

Up to this point, two different types of mental states with a world-to-mind direction of fit and mind-to-world direction of causation have been distinguished: standardly conceived intentions (which are subject to the strong consistency requirement) and proto-intentions (which are subject to a local form of the strong consistency requirement).

6.3 *Proto-intentions are not intentions in action or proximal intentions*

One clarification is now in order. In the literature on action production, occasionally it has been suggested there is a variety of intention that is supposed to trigger and guide the course of the action it represents. Intentions of this variety are known under various names, depending on different conceptions: *intentions in action* (Searle 1983), *proximal intentions* (Mele 1992), *present-directed intentions* (Bratman 1987, Pacherie 2006, 2008). I am here clustering them together in virtue of their functional commonality: that of triggering and guiding the course of action they represent.

An interesting question is whether these states should be considered an additional variety with respect to standardly conceived intentions and proto-intentions. Answering this question relies on taking a stance on an issue that, I believe, so far has not received enough attention: whether and to what extent intentions in action and similar intentions are subject to normative constraints concerning consistency and rationality (cf. Mylopoulos and Pacherie 2017). In the absence of any clarifications on this issue, it is wrong to assume that proto-intentions *just are* intentions in action, present-directed intentions or proximal intentions.

7. *Conclusion*

This article started with the observation that action understanding has mainly been interpreted in terms of two very extreme options: either as involving the ascription of standardly conceived intentions, which would make action understanding a form of full-blown mindreading, or as involving no mental state ascription at all. I have given reasons for considering a middle ground between these two extreme options. Two considerations support the exploration of this middle ground. On

the one hand, one may think that some creatures (e.g., infants of 9 to 18 months of age) *can't* represent standardly conceived intentions. On the other hand, some creatures capable of representing standardly conceived intentions sometimes *needn't* represent them, given the characteristics of the action they are in the position of understanding. Either consideration suffices to consider the following alternative: that action understanding might involve a minimal form of mindreading.

I have presented and explored one way in which action understanding to be minimal mindreading—one that involves the ascription of proto-intentions. By contrast with the posits of Minimal Theory of Mind, proto-intentions are states that one has, and not just states that one ascribes, and they are representations rather than relations.

To sum up, here are the options explored so far concerning how action understanding might occur, together with an indication as to whether they are a form of mindreading and, if so, which form:

In what does action understanding consist?	**Is it mindreading?**
Mere outcome identification	No
Proto-intention ascription	Yes (minimal)
Intention ascription	Yes (full-blown)

The notion of proto-intention can help interpret experiments, such as that by Meltzoff (1995), that have the following characteristics: on the one hand, it would be explanatory advantageous to suppose that the experimental subjects ascribe some form of mental states to the observed individuals, but, on the other hand, we might be reluctant to think of these subjects as mastering cognitively demanding mental state notions, such as the standard one of intention. Proto-intentions, by contrast, are apt to be represented and ascribed more widely across development and species.

References

Andrews, K. 2020. "Naïve normativity: The social foundation of moral cognition." *Journal of the American Philosophical Association* 6 (1): 36–56.

Apperly, I. A. and Butterfill, S. A. 2009. "Do humans have two systems to track beliefs and belief-like states?" *Psychological Review 116* (4): 953.

Behne, T., Carpenter, M., Call, J. and Tomasello, M. 2005. "Unwilling versus unable: infants' understanding of intentional action." *Developmental Psychology* 41 (2): 328–337.

Borg, E. 2007. "If mirror neurons are the answer, what was the question?" *Journal of Consciousness Studies* 14 (8): 5–19.

Bratman, M. 1987. *Intention, plans, and practical reason*. Cambridge: Harvard University Press.

Butterfill, S. A. and Apperly, I. A. 2013. "How to construct a minimal theory of mind." *Mind & Language* 28 (5): 606–637.

Call, J., Hare, B., Carpenter, M. and Tomasello, M. 2004. "'Unwilling' versus 'unable': chimpanzees' understanding of human intentional action." *Developmental Science* 7 (4): 488–498.

Cheney, D. L. and Seyfarth, R. M. 1992. *How Monkeys See the World: Inside the Mind of Another Species*. Chicago: University of Chicago Press.

Csibra, G. and Gergely, G. 1998. "The teleological origins of mentalistic action explanations: A developmental hypothesis." *Developmental Science* 1 (2): 255–259.

Davidson, D. 1963. "Actions, reasons, and causes." *The Journal of Philosophy* 60 (23): 685–700.

Gallagher, S. and Hutto, D. 2008. "Understanding others through primary interaction and narrative practice." *The Shared Mind: Perspectives on Intersubjectivity* 12: 17–38.

Gallese, V. and Goldman, A. I. 1998. "Mirror neurons and the simulation theory of mindreading." *Trends in Cognitive Sciences* 2: 493–501.

Gergely, G. and Csibra, G. 1997. "Teleological reasoning in infancy: The infant's naive theory of rational action: A reply to Premack and Premack." *Cognition* 63 (2): 227–233.

Gergely, G. and Csibra, G. 2003. "Teleological reasoning in infancy: The naıve theory of rational action." *Trends in Cognitive Sciences* 7 (7): 287–292.

Gergely, G., Nádasdy, Z., Csibra, G. and Bíró, S. 1995. "Taking the intentional stance at 12 months of age. "*Cognition* 56 (2): 165–193.

Goldman, A. I. 2006. *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford: Oxford University Press.

Goldman, A.I. 2009. "Mirroring, mindreading, and simulation." In J. A. Pineda (ed.). *Mirror neuron systems: The Role of Mirroring Processes in Social Cognition*. New York: Humana Press, 311–330.

Gopnik, A. and Meltzoff, A. N. 1997. *Words, Thoughts and Theories*. Cambridge: MIT Press.

Gopnik, A. and Wellman, H. M. 1992. "Why the child's theory of mind really is a theory." *Mind & Language* 7 (1–2): 145–171.

Gordon, R. M. 1986. "Folk psychology as simulation." *Mind & Language* 1 (2): 158–171.

Heal, J. 1986. "Replication and Functionalism." In J. Butterfield (ed.). *Language, Mind, and Logic*. Cambridge: Cambridge University Press.

Holton, R. 2009. *Willing, Wanting, Waiting*. Oxford: Oxford University Press.

Hurley, S. 2003. "Animal action in the space of reasons." *Mind & Language* 18 (3): 231–257.

Low, J. and Watts, J. 2013. "Attributing false beliefs about object identity reveals a signature blind spot in humans' efficient mind-reading system." *Psychological Science* 24 (3): 305–311.

Mele, A. R. 1992. *Springs of Action: Understanding Intentional Behavior*. Oxford: Oxford University Press.

Meltzoff, A. N. 1995. Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental psychology*, 31(5), 838–850.

Michael, J. and Christensen, W. 2016. "Flexible goal attribution in early mindreading." *Psychological Review* 123 (2): 219–227.

Montefiore, A. and Noble. D. (eds). 1989. *Goals, No Goals and Own Goals*. Unwin Hyman. Republished by Routledge, 2021.

Mylopoulos, M. and Pacherie, E. 2017. "Intentions and motor representations: The interface challenge." *Review of Philosophy and Psychology* 8 (2): 317–336.

Nanay, B. 2013. *Between Perception and Action*. Oxford: Oxford University Press.

Onishi, K. H. and Baillargeon, R. 2005. "Do 15-month-old infants understand false beliefs?" *Science* 308 (5719): 255–258.

Pacherie, E. 2006. "Towards a dynamic theory of intentions." In S. Pockett, W. P. Banks and Sh. Gallagher (eds.). *Does Consciousness Cause Behavior*. Cambridge, Mass.: MIT Press, 145–167.

Pacherie, E. 2008. "The phenomenology of action: A conceptual framework." *Cognition* 107 (1): 179–217.

Peacocke, C. 2014. *The Mirror of the World: Subjects, Consciousness, and Self Consciousness*. Oxford: Oxford University Press.

Perner, J., and Roessler, J. 2010. "Teleology and causal understanding in childrens' theory of mind." In J. H. Aguilar and A. A. Buckareff (eds.), *Causing Human Action: New Perspectives on the Causal Theory of Action*. Cambridge, Mass.: MIT Press, 199–228.

Searle, J. R. 1983. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press.

Sinigaglia, C. 2008. "Mirror neurons: This is the question." *Journal of Consciousness Studies* 15 (10–1): 70–92.

Sinigaglia, C. and Butterfill, S. 2015. "Motor representation in goal ascription." *Conceptual and Interactive Embodiment. Foundations of Embodied Cognition* 2: 149–164.

Smortchkova, J. 2018. "Seeing goal-directedness: a case for social perception." *The British Journal for the Philosophy of Science*. 71 (3): 855–887

Spaulding, S. 2013. "Mirror neurons and social cognition." *Mind & Language* 28 (2): 233–257.

Stich, S. and Nichols, S. 1992. "Folk psychology: Simulation or tacit theory?" *Mind & Language* 7 (1–2): 35–71.

Tomasello, M. 1999. *The Cultural Origins of Human Cognition*. Cambridge: Harvard University Press.

Tomasello, M. and Call, J. 1997. *Primate Cognition*. New York: Oxford University Press.

Tomasello, M., Call, J. and Hare, B. 2003. "Chimpanzees understand psychological states–the question is which ones and to what extent." *Trends in cognitive sciences* 7 (4): 153–156.

Tomasello, M., Carpenter, M., Call, J., Behne, T. and Moll, H. 2005. "In search of the uniquely human." *Behavioral and Brain Sciences* 28 (5): 721–727.

Whiten, A. 2013. "Humans are not alone in computing how others see the world." *Animal Behaviour* 86 (2): 213–221.

Wilkes, K. 1989. "Explanation – How Not to Miss the Point". In Montefiore and Noble 1989: 194–211.

Croatian Journal of Philosophy
Vol. XXII, No. 66, 2022
https://doi.org/10.52685/cjp.22.66.7
Received: September 1, 2022
Accepted: October 17, 2022

Imagining the Ring of Gyges. The Dual Rationality of Thought-Experimenting

NENAD MIŠČEVIĆ
University of Maribor, Maribor, Slovenia

In her already classical criticism of thought-experimenting, Kathy Wilkes points to superficialities in the most famous moral-political thought-experiments, taking the Ring of Gyges as her central example. Her critics defend the Ring by discussing possible variations in the scenario(s) imagined. I propose here that the debate points to a significant dual structure of thought experiments. Their initial presentation(s) mobilize the immediate, cognitively not very impressive imaginative and reflective efforts both of the proponent and the listener of the proposal. The further debate, like the one exemplified by Wilkes's criticisms and some of the answers, appeals to a deeper, more rational variety of imagination and reasoning. I suggest that this duality is typical for moral and political thought experimenting in general, conjecture that it might be extended to the whole area of thought experimenting.

Keywords: Thought experiment; rationality; imagination; Kathleen Wilkes.

1. *Introduction*

Since the paper is intended as an homage to Kathleen Wilkes, let me start with some memories from Dubrovnik where I met her for the first time and continued hanging around with her each year when I visited Dubrovnik. On each occasion, the two of us have been coming together, endlessly discussing philosophy and enjoying each other's company. In the paper, I shall refer to her as to "Kathy", as we all have been calling her at home in Croatia.

Kathy was the chairman of the executive committee of the Inter-University Centre in Dubrovnik since the mid-eighties, contributing

enormously to the intellectual life in Croatia. Her contribution to analytic philosophy in Croatia and neighboring countries was crucial for the local philosophical development.

But here I want to stress Kathy's incredible wartime solidarity, most clearly manifested in the time of the Serbian army's constant shelling of the town that started in October 1991, culminated a few months later and lasted until May the following year. Kathy was living in Dubrovnik all the time. I remember her from when I came in April 1992, seeing her dressed in Croatian camouflage uniform and passionately commenting on the military situation around Dubrovnik. And she stayed in Dubrovnik after the war ended, helping rebuild Croatian intellectual life. Even later, when I visited her at St. Hilde's college, when her health way deteriorating, she was still dressed in the camouflage uniform, and her favorite topics were her memories from the time of war.

The present paper is dedicated to Kathy's philosophical work, focusing upon the topic to which she dedicated a whole book, her *Real People: Personal Identity without Thought Experiments*, Oxford University Press from 1988. Among other topics, Kathy gives and discusses one example from moral philosophy and this discussion will be the topic of this presentation. But our target is Kathy's criticism of TEs, and we shall be taking her Ring of Gyges example as central.

We first present her stark criticism of this thought experiment and next a defense, due to Cora Diamond (2002), taking her criticism as a paradigm of sophisticated and potentially successful problematizing of intuitions generated by a typical thought experiment (I shall shorten the expression as "TE"). This brings us to the general issue of the source of such debates. We then sketch a general answer, a more systematic sketch, relying on a dual-process account of imagining and reasoning but going further in systematizing the approach specifically in regard to TEs in ethics and political philosophy. Connection with imagination is crucial for our account. We develop the proposal here in two directions: first, connecting issues of imagination to the picture of stages of TE, and second, applying it very briefly to TEs in practical, moral and political philosophy.

2. *The Ring of Gyges – for and against*

As mentioned above, in her (1988) book, Kathy discusses one example from moral philosophy, which will be the topic of this paper. So, here we concentrate on chapter One of her *Real People* book, where the Ring of Gyges TE is presented and criticized.

Here is her announcement:

Examples from philosophy

> We can begin with an example from moral philosophy. As all know, there are several theories about the basis of morality— that it is ultimately for self-interested reasons that we are moral; or that morality derives from

> natural emotions of love, fellow-feeling, generosity, pity, etc.; or that it is based upon rationality; or that it is the result of a fictional social contract; or that it is inevitable, given what we know about sociology and human psychology. (4–5)

And the Ring of Gyges, as presented in Plato's *Republic*, gets in:

> One test suggested to discover the fundamentality of morality is to ask 'what if we all had a Gyges' ring to make us invisible at will?' As we know, no humans are actually invisible, so we cannot try the experiment and see. So we imagine a possible world in which people have such rings, but which is in other respects just like ours. If it seems that in such circumstances nobody would remain moral (i.e. if we think that when we could guarantee getting away with it, we would not bother with moral standards), then, crudely, it looks as though morality is based rather on self-interest than on anything grander. The imaginary state of affairs is the invisibility; one conclusion *may* be that morality must be based ultimately on self-interest. (5)

But then, a few pages later, Kathy offers a harsh criticism of the Ring of Gyges TE, and this will be our focus.[1]

So, let me remind the reader of the basic story of the TE. The story, as told by Glaucon, tells us that the shepherd Gyges discovered one day a big hole in the earth, where he saw surprising things:

> He saw, along with other quite wonderful things about which they tell tales, a hollow bronze horse. It had windows; peeping in, he saw there was a corpse inside that looked larger than human size. It had nothing on except a gold ring on its hand; he slipped it off and went out. When there was the usual gathering of the shepherds to make the monthly report to the king about the flocks, he too came, wearing the Ring. (*The Republic* 359–360, Plato 1991: 37).

And then, a strange thing happened. While he was sitting with the others, Gyges chanced to tum the collet of the Ring to himself, toward the inside of his hand; and when he did this, he became invisible to those sitting by him, and they discussed him as though he were away. He wondered at this, and, fingering the Ring again, he twisted the collet toward the outside; when he had twisted it, he became visible. He tested whether the Ring had this power, and the result was positive. "Aware of this, he immediately contrived to be one of the messengers to the king," the story continues. "When he arrived, he committed adultery with the king's wife and, along with her, set upon the king and killed him. And so he took over the rule" (360 b, 37–8).

Glaucon famously develops the story, turning it into a proper philosophical TE. He invites the reader to imagine that there were two such rings and that the just man would put one on, and the unjust man the other. The result of the imagining is quite shocking "(…) no one, as it would seem, would be so adamant as to stick by justice and bring himself to keep away from what belongs to others and not lay hold of it, although he had license to take what he wanted from the market with-

[1] I hope to address her criticism of personal identity TEs on some other occasion; here we stay with her reading of the Gyges story.

out fear, and to go into houses and have intercourse with whomever he wanted, and to slay or release from bonds whomever he wanted, and to do other things as an equal to a god among humans" (360 b). In so doing, Glaucon continues, one would act no differently from the other, but both would go the same way. "And yet, someone could say that this is a great proof (*mega tekmerion*) that no one is willingly just (*hoti oudeis hekon dikaios*), but only when compelled to be so" (360 b). And he concludes that there is no deep difference between the just and the unjust man; in real life people act justly merely because of fear of punishment; "it looks as though morality is based rather on self-interest than on anything grander," as Kathy puts it (5) This is what the TE clearly suggests.[2]

Now, the main point of Kathy's criticism is that the TE is superficial; and she develops her accusation of superficiality in a most interesting way. This will be our topic here, so we shall start by quoting her extensively:

> Consider ...Gyges' Ring: before we can make sense of this thought experiment, several points press to be answered— there are relevant background conditions that need to be known before we can draw any conclusion(s) from the imagined phenomenon. We need more information than we yet have about this 'possible world'. (11)

And now Kathy comes up with her main line of criticism. What exactly can Gyges do, we are invited to ask.

> For instance, is the owner of the Ring to be intangible as well as invisible? That makes a substantial difference to the issue at issue: if he is not intangible, he might by mistake bump up against, and get arrested by, a policeman, or get his hand slammed shut the till-drawer. Thus, a potential criminal may yet have self-interested reasons for staying within the bounds of morality. (11)

Things get worse for Glaucon. Here is her further criticism:

> Is there anything that would count as 'punishment' for an invisible and intangible agent? If so, what—and how unpleasant would it be? If you are both invisible and intangible, could prison walls hold you? And if they could not, could you hold a gun, or a caseful of banknotes? Again, would others know that one owned such a ring? If so, then there might be extra reasons for remaining moral: viz., that unsolved crimes might otherwise be ascribed to you. The point is that the purpose of the thought experiment cannot be met unless such questions are answered: they are deeply relevant. The background is inadequately described, and the results therefore inconclusive. (11)

The criticism is quite sharp, and it leaves for us no morals of the TE. No wonder, critics reacted. Here we shall concentrate of the answer offered by Cora Diamond in her (2002) paper. Talking of Gyges she says: "The objection seems to me to miss its mark" (231). She usefully summarizes

[2] I am thankful to Boris Vezjak for critically discussing my paper, in seminars and in his chapter in my Festschrift, Vezjak (2017). Thanks also go to Miomir Matulović for his detailed critical discussion.

Kathy's methodology. According to her reading the underlying idea of the criticism is that, if thought-experiments can be fruitful in philosophy, their fruitfulness will be dependent on their having a determinate outcome, "like thought-experiments in physics, (…) which have an outcome determined jointly by the conditions described together with background conditions" (231). In such successful experiments, we know what factors are being juggled; "for the rest of the natural world as we know it is in place and has to be for the experiment really to have a determinate result, for it to be fruitful and able properly to convince us of something" (231).

However, this demand is irrelevant for Glaucon's argument. All he needs "is a thinking away of the probabilities large and small of discover that might attend unjust action. He does not have to provide the details of the imagined natural laws of a world in which some individuals would be able to perform unjust actions with confidence in not being discovered" (232). This, she says, offers us an idealized version of something we know to happen, namely that confidence in not being discovered is frequently an element in people's deciding to act in ways considered unjust.

This brings us to the question that will take us to the central issues to be discussed here. Where does the conflict between the two versions, Kathy's and Diamond's, come from? It looks like the discussion offers a two-stage scenario:

First, the crucial, immediate stage and the kind of imagining that accompanies it, that is the first, spontaneous reaction and answer: If I were sure I cannot be discovered, I would steel, and murder and rape! And this supports an immediate general stance: confidence in not being discovered is crucial in one's decision to act unjustly.

Second: the stance is taken by the interlocutor to open space for a deeper philosophical discussion, of the kind quite different from the quick presentation starting the dialogue.

We shall be looking at this structure throughout the rest of the paper. Note that the phenomenon clearly generalizes to other theories mentioned by Kathy in the text. For instance, to the whole wide and crucial important genus of contractualist political TEs, mentioned by Kathy. She notes that, as all know, there are several theories about the basis of morality, and one option she mentions is "that it is the result of a fictional social contract" (6). Take the version proposed by Habermas:

> Each of us must be able to put themselves into the position of all those who would be affected by the performance of a problematic action or the adoption of a questionable norm. (1993: 49)

What is assumed are the willingness to communicate, rationality and full information on the side of the interlocutor. But then, at a late, reflective stage of the TE, a counterpart of Kathy can come problematize the assumption: What if the agent is not willing to communicate? Does she lose her moral status?

Similarly with other idealizations assumed by contractualist philosophers. For instance, Scanlon famously talks about reasonableness. For him, it is the ability of perspective taking that is crucial: I have to think of reasons that the person I am interacting with cannot reasonably reject ("...an act is right if and only if it is justifiable to others on terms they could not reasonably reject" 1998: 189).

The discussion of such a quasi-idealization does not belong to the immediate, non-reflexive imagining; our hypothesis is that it is a matter of later, reflexive stages. The same with other TEs in practical philosophy.

Take the Original Position TE due to Rawls. The reader is asked to imagine s/he is free of envy, and this is crucial for the TE. But can one really do it? Can I be happy with the imagined situation where my neighbor is ten times more talented and three times richer than I am? Well, it's just idealization! But is it an acceptable one, Kathy's counterpart would ask.

So, let me generalize. The standard form of debate in practical philosophy (also wider) concerning TEs, from Plato on: the proponent, says Plato, presents a simple scenario (like in Kathy's example the Ring of Gyges). He raises one or two crucial questions, and presses the interlocutor for an answer. The interlocutor reflects very briefly, and comes with a short answer. The answer is normally taken by the proponent to suggest a view, even a philosophical one, and the proponent develops it into a sketch of a theory.

Typically, a further discussion starts and continues, for instance, with the Ring of Gyges, for two millennia and a half. In the discussion, the critics point to holes in the original story, suggest accounts alternative to the originally proposed one, and the debate goes on, endlessly.

3. *The dual structure of thought-experiments*

Here is then the crucial question: what is it about TEs that supports the endless number of cases like the ones mentioned? This is our main question to be discussed in the sequel. Let me illustrate it in a bit more detail, going back to Gyges and his Ring. First, imaginative reasoning. Remember the proponent suggesting a scenario and raising his question: "Imagine yourself being invisible. And facing a large quantity of money. What would you do?" The interlocutor replies that, of course, he would take it and run away. "What about attractive young women around?" and so on. And the general conclusion follows.

What would a cognitive psychologist say? She would ask us to note that the subject didn't think of further alternatives. He is invisible, but he remains tangible; otherwise he could not take money, or harass the attractive young women in the story. She would point to us that we imagined and reasoned on the basis of information directly available, ignoring the slightly more distant option. This is called availability heuristics (see Tversky and Kahneman 1973).

When you are challenged, you might start thinking of the other option. But this then is not quick heuristics, but a more reflective imagining and reasoning. The contrast is nicely captured by Michael T. Stuart who actually proposed the two terms in his (2021) paper. He writes:

> Call "imagination$_1$" the unconscious, uncontrolled, effortless cognitive interaction with objects not currently present to sensory experience.
> Call "imagination$_2$" the controlled, effortful and conscious cognitive interaction with objects not currently present to sensory experience, again. (1337)

Applied to reasoning, we have reasoning$_1$ (imaginative or otherwise) and reasoning$_2$ (imaginative or otherwise).

Using Stuart's terminology, the psychologist can then suggest that the initial reasoning, with the proponent asking the interlocutor to imagine oneself being invisible and facing a large quantity of money involves imaginative reasoning$_1$. Now, the issues that come up after some reflection, like whether the person is also intangible, and if yes, she can do nothing with her hands, so they don't interact with objects in her surrounding, demand imaginative reasoning$_2$. Glaucon exemplifies imaginative reasoning$_1$, while Kathy and Diamond exemplify imaginative reasoning$_2$. This brings us to a more general cognitive account.

Our topic is now the contrast between immediate reactions (like in Glaucon and his intended reader) and the protracted later debate (Kathy and Diamond style). What kinds of imagining and reasoning are involved? We suggested the contrast between two kinds of imagination, borrowing from Stuart the contrast between imagination$_1$ and imagination$_2$. Cognitive psychologists talk about system-1 and system-2 functioning.

Let us apply the distinction to ethical TEs. Think in terms of stages. The standard form of debate in practical philosophy (also wider) concerning TEs, from Plato on suggests the following stages:

First, the stages of the use of imagination$_1$: The presentation of the scenario to the experimental subject (either the author of the scenario herself, or an interlocutor), the (typically imaginative) contemplation of the scenario and some, let us say minimal, piece of reasoning, and finally the decision ("intuition") concerning the thesis/theory to be tested. The proponent presents a simple scenario. He raises one or two crucial questions and presses the interlocutor for an answer.

The interlocutor reflects very briefly and comes up with a short answer. The proponent takes typically the answer to suggest a view, a philosophical one, and the proponent develops it into a sketch of a theory.

Next, the stages that demand more sophisticated discussion. When one is challenged, one might start thinking of options not mentioned in the initial scenario. But this thinking is not following quick heuristics, but a more reflective imagining and reasoning, the one we marked as reasoning$_2$ and imagination$_2$. Typically, in successful cases, such a further discussion starts and continues, for instance, with the Ring of Gyges, for two millennia and a half. In the discussion, the critics point to

holes in the original story, suggest accounts alternative to the originally proposed one, and the debate goes on, endlessly. Often, the scenario is varied, and subject is invited to draw the conclusion from a series of answers to a series of varied scenarios; one can use the term "intuitive induction" for this procedure. Normally, the conclusion is then compared and possibly contrasted to the dominant views in the field, from commonsensical to theoretical, scientific or philosophical ones. If all goes well a reflective equilibrium is reached. So much about TEs of the sort we mentioned, from Gyges to contractualism.

A lot of work should be done to generalize it further.

Now, what is it about TEs that supports the endless number of such cases? This is our question here. In more recent literature, some dispersed fragments of an answer have been given, by prominent theoreticians in the field, like Stuart (2021), Goldman and Jordan (2013) and Saunders (2009)

Goldman and Jordan (2013) focus on one aspect, mindreading, and one method, simulation (in the Gyges example this would apply to Gyges' understanding of the king, the queen, the guards and so on). They also distinguish two levels:

1. Low-level simulational mindreading, e.g. emotion *mirroring,* and
2. High-level simulational mindreading (Goldman and Jordan 2013: Sections 3 and 4).

Several other authors are reflecting in the similar direction. For instance Saunders in his 2009 paper with a telling title "Reason and intuition in the moral life: A dual-process account of moral justification" suggests that understanding of moral intuitions "requires appealing to a dual-process view of moral judgement that regards moral intuitions and moral theories as belonging to different mental systems" (2009: 335).[3] And he points to the connection with duality of cognitive systems: "We can think of moral intuitions, like any other kind of intuition, as System 1 judgments, and consciously and explicitly developed moral theories can be thought of as the outcomes of System 2 processes" (2009: 340).

We suggest that this duality should be applied to the understanding of TEs in practical philosophy, i.e. to moral and political TEs. A lot of work should be done to fully generalize it.

Back to the imagination in philosophy. Here is a further illustration, this time not a contractualist one. Here is the famous "Trolley problem" due to Philippa Foot and formulated in 1967. We shall quote the simple formulation due to Judith Jarvis Thompson, who made a significant contribution to the discussion of the TE:

> Suppose you are the driver of a trolley. The trolley rounds a bend, and there come into view ahead five track workmen, who have been repairing the track. The track goes through a bit of a valley at that point, and the sides are steep, so you must stop the trolley if you are to avoid running the five

[3] See also the section Two systems and the possibility of reflective equilibrium.

> men down. You step on the brakes, but alas they don't work. Now you suddenly see a spur of track leading off to the right. You can turn the trolley onto it, and thus save the five men on the straight track ahead. Unfortunately, there is one track workman on that spur of track. He can no more get off the track in time than the five can, so you will kill him if you turn the trolley onto him. (1985: 1395)

Here are the stages: We begin with stage one, the question being asked. Stage two, the question is understood by the subject. Stage three offers the tentative conscious production, say building the picture of the two tracks with workers, all done at a conscious level. Stage three brings additional unconscious production and is probably controlled by the relevant competence at the unconscious level (some geometry and commonsense physics might be needed to imagine the scenario in sufficient detail). At stage four, the subject arrives at the immediate, spontaneous verdict, often non-conscious, for instance, "Yes, I would turn the lever and save the five men." One might think of an additional stage of subpersonal empirical theorizing by Central Processing Unit. (I, the reader, might imagine workers from abroad, say Mexicans, I might imagine young and healthy workers, or older and tired ones, and so on, all motivated by my views on the working class and the like.) At stage 5 comes the immediate spontaneous answer (intuition): "Yes, I would turn the lever and save the five men." At stage six we have varying and generalizing, intuitive induction at both conscious and unconscious levels. For instance, it seems to me in the trolley case, that I would turn the lever and thus save five by sacrificing two; but what if the two are (a) small children, (b) very talented artists, (c) my friends? Stage seven offers the general belief, for instance that I would turn the lever no matter what.

We might think of a further stage in which I wonder how the result, the general belief, fits with my other considered judgments (intuitions), with theories I believe in and so on? For instance, I would turn the lever since I think five lives are more valuable than two, no matter whose lives the latter are. But why do you believe this? Why don't you give additional weight to children, since they have more time left to enjoy their lives? Because I think the value of each life is the same as of any other. And so on. If I arrive at a satisfactory view of the whole, this will yield a "reflective equilibrium" at conscious level in which my views form an equilibrated structure.

But what about reasoning with imagination (see Myers 2021)? In TEs reasoning goes with imagination. What is specific for it? My proposal is the following: to each kind of imagination we should join the corresponding kind of reasoning:

$$\text{imagination}_1 \rightarrow \text{reasoning}_1 = \text{imaginative reasoning}_1$$
$$\text{imagination}_2 \rightarrow \text{reasoning}_2 = \text{imaginative reasoning}_2$$

The picture is crucially important for evaluating the rationality of TE-ing and its normative status, say in terms of epistemic virtue vs. epistemic vice.

Similar features of System1 processing and of imagination$_1$ are easily recognizable in everyday reasoning. These days, in times of Ukrainian war, you ask an ordinary person: Would you accept refugees? The typical interlocutor can think of two contrasting pictures in his mind, depicting Ukrainian vs. Arab refugees. The Ukrainians are women, Christian, attractive (in the case of my country, Croatia, they also speak a rather similar language). Arabs are typically imagined as men, they are Muslims, mostly young ones (and they speak an incomprehensible language). Here, the heuristics of stereotyping is powerful and omnipresent; the fact that many Arab refugees are women is simply forgotten, and so on; stereotyping insists on contrast:

- Stereotyping
  Ukrainian vs. Arab

And imagination$_1$ works intensely, accompanied by the reasoning of the same kind. The conclusion is clear: "Ukrainian refugees are highly acceptable, Arabs should be rejected in any case," says our interlocutor.

One can talk of minimal rationality in the case of the use of imagination$_1$. And of fuller rationality of the use of imagination$_2$. Similarly, one can note a minimally virtuous status of the use of imagination$_1$ and epistemically virtuous status of the use of imagination$_2$. The contrast has been studied by various authors, psychologists and philosophers (see e.g. Kung 2016).

4. *Conclusion*

We noted that Kathy Wilkes has been pointing to superficialities in the most famous moral-political TEs, taking the Ring of Gyges as her central example. Her critics defend the Ring, by discussing possible variations in the scenario(s) imagined. I have been arguing in the paper that the debate points to a significant dual structure of TEs, of the kind anticipated by Stuart (2021). The central TEs in practical philosophy requires a several-stage work by interlocutors: most importantly, an early stage culminating in an intuitive answer, crucial for the TE, and later stages of doubts, debate and reflective equilibrating.

The initial presentation(s) mobilize the immediate, cognitively not very impressive imaginative$_1$ and reflective$_1$ efforts both of the proponent and the listener of the proposal. The further debate, like the one exemplified by Wilkes's criticisms and some of the answers, appeals to a deeper, more rational variety of imagination and reasoning, imagining$_2$ and reasoning$_2$. The pessimists, most prominently Kathy Wilkes, famously concentrate on weaknesses of the intuitive answer, suggesting that no further elaboration can help with them.

I suggest that this duality is typical for moral and political thought experimenting in general, and conjecture that it might be extended to the whole area of thought experimenting. And I suggest that there is a rationale for a more optimistic reading of practical TEs, grounded on the standard cognitive account of ordinary imagination-and-reasoning.

The picture is crucially important for evaluating the rationality of TE-ing, and of its normative status: epistemic virtue vs. epistemic vice. One can talk of minimal rationality of the use of imagination$_1$ and fuller rationality of the use of imagination$_2$. Similarly, with minimally virtuous status of the use of imagination$_1$ and epistemically virtuous status of the use of imagination$_2$.

The division between early, intuitional, and later, reflective stages, thus mirrors the dual nature of normal human processes of imagining and reasoning. This has been noted in the literature but without a clear connection with the duality of stages, sometimes noted, but not made explicit. We argued that the early stages/later stages division roughly corresponds to the division between system 1 and system 2 imagining-reasoning.

How should the friends of imagination reply? We need mechanisms for self-improvement, in order to have workable TEs, they can note. The attention to imagination can help solve some recurring problems in the debate (and in the meta-theory) of TEs, as we have argued above. Kathy's criticisms suggest the direction to take. The weaknesses of the early intuitional stages are natural consequence of the limited rationality of system-1 cognition, and are routinely ameliorated in the later stages, exhibiting system-2 reflection.

The optimist wins: the job of philosophy is to guide us from the spontaneous but superficial system-1 reasoning and imagining to reflective, system-2, epistemically virtuous elaborations. And TEs are natural, almost ideal means for achieving this. This also explains their omnipresence in philosophy and the rich and varied millennial history of their most famous instances.

No wonder Kathy dedicated so much attention to them, and we should follow her in this! So, let this paper be a homage to Kathy and to her philosophical insight!

References

Diamond, C. 2002. "What if x isn't the number of sheep. Wittgenstein and Thought Experiments in Ethics." *Philosophical Papers* 31 (3): 227–250.

Goldman, A. I. and Jordan, L. 2013. "Mindreading by Simulation: The Roles of Imagination and Mirroring." In S. Baron-Cohen, M. Lombardo and H. Tager-Flusberg (eds.). *Understanding Other Minds: Perspectives from Developmental Social Neuroscience*. Oxford: Oxford University Press.

Habermas, J. 1993. *Justification and Application*. Cambridge: Polity Press.

Kung, P. 2016. "Thought Experiments in Ethics." In A. Kind and P. Kung (eds.). *Knowledge Through Imagination*. Oxford: Oxford University Pres, 227–245.

Matravers M. (ed.). 2003. *Scanlon's Contractualism: Readings and Responses*. Frank Cass Publisher.

Myers, J. 2021. "Reasoning with Imagination." In C. Badura and A. Kind (eds.). *The Epistemic Uses of Imagination*. London: Routledge, 103–121.

Saunders, L. F. 2009. "Reason and intuition in the moral life: A dual-process account of moral justification." In J. Evans and K. Frankish (eds.). *Two Minds: Dual Processes and Beyond*. Oxford: Oxford University Press, 335–354.

Scanlon, T. M. 1998. *What We Owe to Each Other*. Cambridge: The Belknap Press of Harvard University Press.

Stuart, T. M. 2017. "Imagination. A Sine Qua Non of Science." *Croatian Journal of Philosophy* 17 (1): 9–32.

Stuart, T. M. 2021. "Towards a dual process epistemology of imagination." *Synthese* 198: 1329–1350.

Tversky, A. and Kahneman, D. 1973. "Availability: A Heuristic for Judging Frequency and Probability." *Cognitive Psychology* 5 (2): 207–23.

Vezjak, B. 2017. "The Ring of Gyges and the Philosophical Imagination." In B. Borstner and S. Gartner (eds.). *Thought Experiments between Nature and Society: A Festschrift for Nenad Miščević*. Cambridge Scholars Publishing, 410–424.

Wilkes, K. V. 1988. *Real People. Personal Identity without Thought Experiments*. Oxford: Clarendon Press.

Croatian Journal of Philosophy
Vol. XXII, No. 66, 2022
https://doi.org/10.52685/cjp.22.66.8
Received: July 30, 2022
Accepted: September 17, 2022

Purposiveness of Human Behavior. Integrating Behaviorist and Cognitivist Processes/Models

CRISTIANO CASTELFRANCHI
National Research Council, Rome, Italy

We try not just to reconcile but to "integrate" Cognitivism and Behaviorism by a theory of different forms of purposiveness in behavior and mind. This also implies a criticism of the Dual System theory and a claim on the strong interaction and integration of Sist1 (automatic) and Sist2 (deliberative), based on reasons, preferences, and decisions. We present a theory of different kinds of teleology. Mere "functions" of the behavior: finalism not represented in the mind of the agent, not "regulating" the behavior. Two kinds of teleological mental representations: true "Goals" in control-theory, cybernetic view, with "goal-driven" behavior (intentional action); vs. Expectations in Anticipatory Classifiers: a reactive but anticipatory device, explaining the "instrumental" (finalistic) nature of Skinner's reinforcement learning. We present different kinds of Goals and goal processing and on this ground the theory of what "intentions" are. On such basis, we can discuss Kathy Wilkes's hint about the necessarily linguistic formulation of "intentions"; with the hypothesis that her intuition is not correct for any kind on "intention" which may be represented in sensory-motor format, but correct for "volition" and our will-strength for socially influencing ourselves.

Keywords: Teleology; goal theory; intentions; behaviorism; dual System.

1. *Premise: Claims and Moves**

The *claims* are the following ones:

It is time—also thanks to the pressure due to the neuro-foundation of psychological models—to reconcile Cognitivism with Behaviorism (two philosophical and historical enemies). Not just to reconcile but to "integrate" them, by not simply explaining coexistence of postulated mechanisms but their systemic interaction and interference. This attempt will in part overlap with a reunification of System 1 and System 2 postulated in the "Dual System" view of the mind.

Main moves necessary for this integrated theory in fact are:

– A critical revision of "dual process" theory:[1]

(i) It assembles as a unified "process" (automatic, fast, associative, holistic) several very different mechanisms; or just opposes "affect" and "reason" (Loewenstein and O'Donoghue 2004)

(ii) These ("multiple" not "dual") processes do not just compete and prevail one on the other, but interact and cooperate (for example, in the complex and hybrid "value" of a goal, both belief-based, reasoned, and just "felt" ("somatic markers", etc.).

– Making formally clear the fundamental distinction between the *two kinds of finality*, of "goal", impinging on animal behavior: *mental goals* (based on control theory models), vs. *external goals, mere "functions"* (based on selection processes). A frequent mistake of psychologists (Castelfranchi,1999) is to interpret any clear purposefulness of human behavior in terms of conscious or unconscious intentions in the mind of the individual (Bargh et al. 2001).

– In this frame, we need—as said—a more "representational view" of conditioning.[2] However, in the "mentally represented" teleological devices we will distinguish true "Goals" from Expected Results reinforcing and explaining that conduct. It is crucial to make clear the difference between these *two kinds of anticipatory representation* governing the action. And modeling on such basis the "instrumental" (finalistic) nature of Skinner's conditioning.

– Modeling the layered *integration* of reactive/automatic devices and of intentional and reasoned actions; for example, by implementing higher level deliberated action in underlying automatic classifiers.

One should also try to:

– Explain how conditioning, reinforcement learning (both Pavlovian and instrumental), act also on symbolic "mental representations" pos-

* This is more a palimpsest of a work in progress than a balanced paper. It contains a vision and some basic claims; a schema of the main moves that should be done; and exploration of a few specific issues including an homage to Kathy Wilkes' intuitions.

[1] Nowadays very popular. Literature is very broad and with different positions (Caccioppo, Kahneman, Sloman).

[2] And putting aside some really reductive proposal of behaviorism, like the reduction of guilt feeling to worry for punishment!

tulated by Cognitivism (beliefs, expectations, goals...), and interfere (not only compete) with the high-level cognitive processes.
– To discuss the notion of "reward" and its function, and to put aside "hedonism" (pleasure) as the unifying motivation.

Let us be a bit more analytical on some of those issues. At the end, on the basis of theory and modeling of "intention", we will discuss an interesting thesis of Kathy Wilkes, as an homage to her deep thinking.

2. *The anticipatory nature of the mind: two devices*

It is very important to understand the anticipatory nature/origin of mind (and the more general "augmented reality" function of the brain) and the creation of "endogenous" representations/worlds: not output of current perception input, but self-generated by memory activation, generative recombination, imagination and simulation. A fictional world where to act, learn, solve problems.

However, we have to distinguish two very different anticipatory devices: Anticipatory Classifiers (ACs) (bottom-up, responsive) versus true goals (control theory, top-down) (Pezzulo et al. 2008). In both cases, there are "expectations" but with different roles: in ACs just reinforcement function; in Goals cybernetic set-points, monitoring and adjusting, (sub)planning. In both cases, there is "failure" (frustration) or "success."

ACs are very important for contrasting a primitive behaviorist, conditioning-based explanations of some behaviors just in terms of S-R, Condition-Action ("production rules" or Classifiers) models of reinforcement learning.

However, we also have to be reminded that there is another kind of finalism in animal and human conduct, not represented at all: mere "functions" of that behavior (or feature).

2.1. *Mere "functions" as not mental/represented goals*

As already said there are *two kinds of teleology*: (i) mentally represented (and eventually intended) or *psychological goals* that regulate our conduct; and (ii) non-mental goals, just emergent and self-organizing "functions" (social or biological) impinging on our individual and collective behaviors. Let us use the term "goal" just for the internal control system, the mentally represented objective; and the term "function" for the external selecting finality of a feature or a behavior (Conte 1995; Castelfranchi 2001).

Behavioral functions are simply effects of behavior, which give a positive feedback on it, reinforce or select it, and reproduce it. *Functional effects*, usually unintended (desirable or even undesirable) and not understood, but such that they have feedback and select that behavior or entity.
– *Selective/evolutionary "functions"* of behaviors or features (not only of behaviors; also the features of living being have a function: an adaptive effect)

There also are:
– *Technical functions:* objects also have finalisms: they are "made *for*" and "used *for*": function of the object / tool.

Also in cognitive intentional Agent there can also be merely "function" governed conducts: effects of behavior which go beyond the intended effects but which can successfully be reproduced because they reinforce the agent's beliefs and goals that give rise to that behavior.

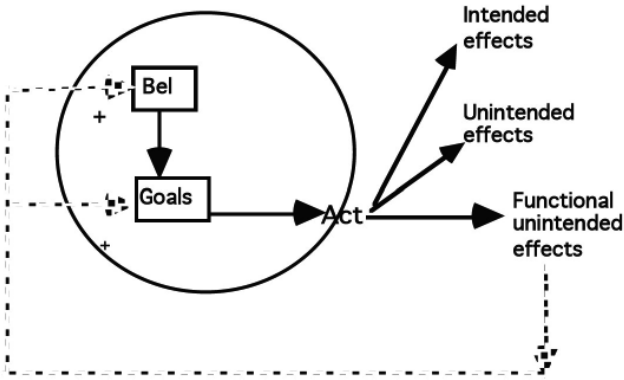

Fig.1 Functional unintended effects

2.2. *Teleology in Dual Processing System 1*

Do we intend all the goals/finalities of our behavior? We do not "intend" all that we "pursue" ("functions"). Are all the expected positive results, the achieved goals "intended" results? No, we do not "intend" all we expect. As presented in this frame, we need a more "representational view" of conditioning first of all by making clear the difference between two kinds of *anticipatory representation* governing the action: true "goals" for goal-directed action vs. "anticipatory classifiers"—as special kind of "classifiers".[3]

The format of *Anticipatory Classifiers* is: C → A + Exp

Matching Condition activates an Action + Expectation (anticipated results).[4]

Similar to intentions but not intentions: not a "goal-driven" behavior whose model is TOTE model of Miller, Galanter, Pribram (1968) characterized by a top-down processing (from the goal to the action) not a bottom-up process:

[3] They are "Classifiers" (Cond ⇒ Action, S-R like) but they are based on *Anticipatory Representation*, on Expectations. Condition → Action + Exp. And their reinforcement is *due to the confirmation of the expected result* (Exp) (Pezzulo et. al. 2008).

[4] Moreover, Exp, or anticipated representation of perceptual nature, is *an expected sensation that* determines the "success" or failure of the act. Sensation that might also be *proprioceptive* or *enteroceptive*, that is, about a bodily state: a "feeling'.

First comparing the GOAL (starting point) against the World, and then (in case of mismatch) searching for/activating an action.

This kind of *Proto-Goals* [5] (Exp in ACs) and *proto-intentional conducts* are important in human agents for several reasons/functions:

- evolutionary and developmental stages;
- coexistence of different *teleonomic* mechanisms (not simple S-R) that govern and contend for behavior;
- Routine and automatic components of conduct; also of the intentional conduct;
- For explanation of—in our view—"Instrumental or Operant" conditioning and learning (Skinner), and why it is *seemingly intentional;*
- Probably also for explaining the "reinforcement learning" component of *neurotic persistence*, and its *circularity* in particular, when combined with the idea of sensorial and especially entero-ceptive expectations (feelings), sensations from/about my own body ex."relief" as a reinforcing—non realized—experience/feeling (ex. social anxiety; avoidance).

Moreover this mechanism and this anticipated representation and expectation is not necessarily conscious. The subject can be unaware of it, and this kind of primitive "control" can be merely "automatic" (like using the brakes and expecting the car to slow down) (Castelfranchi 2001).

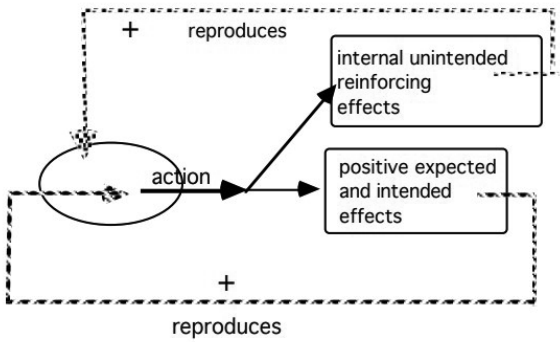

Fig. 2 Unintended effects reinforcing the conduct

In our view, for example, ACs are crucial for explaining the "reinforcement learning" component of *neurotic persistence*, and its *circularity*. We wonder (but I am not a clinical psychologist!) if this dynamics is underlying "akrasia" experience in general. When I act in conflict with my best preference, what I think of would be better to do. We do not think that such a conflict and scission is just a conflict between affective im-

[5] "Proto" because they are similar to but not true goals, but also because reasonably they were the first form of mentally represented results, anticipated, and finalizing the conduct; before true goals and intentional actions.

pulses versus reasoned planning (like in Lowenstein version of "dual processing" (Loewenstein and O'Donoghue 2004)). Nor do we think—see below—that this is the result of a double reasoned decision process were there are *consciously* calculated advantages but also (prevailing) *unconsciously* calculated "secondary" outcomes with greater utility. Are neurotics perfectly *crypto-rational decision makers*? We guess that it is a matter of a conflict between a merely conditioned activated conduct vs. an intention-driven attempt.

How many human conducts are read as strictly goal-driven (intentional, preferred) while they are just conditioned?

2.3. *"Secondary advantage"*

In our view (Castelfranchi 1998, 1999) "secondary advantage" exists and operates, but it is not a "calculated" advantage we put in our reasoned decision, and we "rationally" *decide* for it but unconsciously (against what we consciously believe to prefer and would like to do). We are not rational but *unconscious decision makers*. The behavioral output is not the outcome of a reasoned evaluation of pros and cons; we do not choose what we consider better for us. The underlying model is a different one; it is a DUAL processing model, where two systems (the automatic, nonintentional reinforcement basic one, and unconscious and the deliberate one) compete with one the other, and the system based on "instrumental" reinforcement learning and on anticipation (but not "intention"!) of the reward can win, and we do something different (perhaps even do not really understanding "why") from what we would reasonably prefer. And we perhaps find some post-hoc and ad hoc explanations (reasons) of our choice, not necessarily the right ones! We expect a reward and act "in order" to obtain such (internal) reward (pleasure, pain avoidance, relax, stop anxiety…) but such an expectation is not our "goal" in control theory and psychological sense. It is just the Exp of an anticipatory classifier, maintained/reinforced by its activation, execution, and success/confirmation of the result. We are forced by such reactive and sensation-based but prospective device. And we can in fact also feel "without control and real decision," acting against our good and intention, coherced.

By analogy it is not true that we usually intentionally try to avoid to elicit a bad impression "in order" *not to experience* the unpleasant feeling of shame; we want a good reputation and esteem: this is our motivation. It is false that we avoid to do something bad and unfair "in order" *not to experience* the uncomfortable guilt feeling; we want not to be bad, but to be correct and moral. However, the avoidance of such unpleasant feeling states is there; it possibly is a negative reinforcement of certain actions and, in a sense, our behavior "in order to" avoid them, has such a finality; but it is not—usually—our aim.

3. *Reinforcement Effects on Cognitive Representations*

The other fundamental unification move of behavioristic models and devices and cognitive mental "representations" and processing, is not just to put the two systems in competition or in convergence one with the other, but to say that behavioristic rules also apply to higher level cognitive mental representations and not just to perceptive stimuli and pre-planned executive responses (Bargh and Ferguson 2000; Castelfranchi 2001).

For example, it plays a very crucial cognitive role in the fact that we act on the basis of what we believe, but many of these beliefs are not explicitly formulated or activated, taken into account, and reasoned about. However, they are not challenged ("surprise"), they remain just *presupposed*. There are a lot of "presupposed tacit assumptions" under any action of ours. For example, when I decide to walk in that direction (to go to my office) as usual and routine-like way, not only that I implicitly believed that my office was there (since this was at the beginning—before building a mere routine), but I also "assume" that the floor will support me, that it is safe. I have no reason for thinking about that (consciously or unconsciously), such assumption is not active at all. However, even these presupposed and implicit assumptions (which can also be formulated in a not propositional format i. e. sensory-motor or procedural) if the action succeeds, they get an automatic feedback of confirmation, they are more stable, reinforced ("credible"), and remain presupposed. This also is one of the reasons why failure is a crucial experience for discovering, understanding, and learning.

This *doxastic reinforcement,* the unconscious mechanism is so important in human cognition that it was the advice of Pascal about how to arrive to believe something you cannot rationally believe: you have to act "as if" you believe it, "as if" it was true that… and you will come to believe so. And it is also a classical prescription of cognitive-behavioral psychotherapy in order to abandon some dysfunctional (for Beck[6] "irrational") belief you have: recognize that you can change your mind; *"stop acting or thinking on the basis of the old belief"*, and act in the light of a new belief, and continue to behave in the new way even though it feels phony to act so, and "*that will cause the new belief to become real and a part of your 'natural' behavior".* This reinforcement effect due to the feedback of a successful action does not only apply to the background (implicit or explicit) beliefs, but also to the adopted plan and means (and to beliefs that are valid), to the goal (by increasing its value as for its attainability and probability). It also reinforces our attachment to our final motivating goals and to our values. Not by reasoned conclusions, evaluations, meta-beliefs but by some sort of "reward" to our assumption, planning, objectives, choices, etc. For example, a successful "action schema" increases—by feedback—its accessibility and affor-

[6] Aaron Beck, the father of "cognitive behavioral therapy."

dance, the probability to be retrieved and chosen next time and some sort of index/measure of its validation. This feedback reinforcement is the fundamental route for their automatization, packing, routinization and habits construction.

A different case is "affect/feeling as information." The normal, canonical cognitivist view is that the cognitive appraisal (beliefs, evaluations) of an event is the forerunner of the emotional response; however, the other way around also exists: feeling something as evidence, as base for believing something. For example, feeling some worry, fear, as a base for *believing* that a threat, some danger is there. Now, given this reverse process the two mechanisms can be combined in a vicious circle (like in panic crisis):

Bel: "There is danger!" ⇒ "Fear" ⇒ feedback *reinforcing* the belief of danger.

4. *Reconciling System 1 and System 2*

First, the conflict between Syst1 and Syst2 is not a matter of a conflict between "rational" or "cultural" aspects against "instinctual" aspects (in case between mere learning by reinforcement vs. true resolutions). Nor is it simply a matter of a conflict between "rational" mechanisms against emotional mechanisms (like in Loewenstein's model).

(i) System 2 is "*reason-based*," that is based on beliefs and evaluations, but this is different from "rational."
(ii) Moreover, both Intentions and activated Classifiers can have *an emotional-impulsive origin*.

Second, the two systems are not just in competition and conflict.[7]

Syst1—it's true—can bypass *deliberation* at all; they compete with each other. Not only "decision" produces action, but also other mechanisms that bypass a real deliberation process:

- reactivity and rule-based behavior
- emotion impulses (like in Lowenstein's view)
- habits and script-based behavior; routines, practices and conformity

But this is not the full story.

Syst1 (with its intuitive, impulsive "values" and "reasons" for preference ("reasons of the heart") and Syst2 (with its reasoned, arguable evaluations and preferences) can interact/interfere with each other, and we can decide by taking into account both: the reasoned values (the reason of the Reason) and the felt values (the heart's reasons) (Castelfranchi 2016). Moreover, Syst1 and Syst2 can be translated one into the other. Many acts originally "driven" by intentions, "in view of," etc. can *become automatic*: routines, habitual, reflex-like, respondent. Classifiers activated by conditions and context, where the original "purpose"

[7] For a deep criticism of the *duality of mind* see also Viale (2019).

remains inactive and implicit. Example: "automatically stop at the red light" that originally when we were learning to drive the car was a real decision. On the other hand, a merely automatic reaction can become problematic, not executable in a given context and we have to reformulate an intention and make a real decision; like at red traffic light but with the siren of an ambulance behind us.

However, the most important form of interaction (not separation) of the two systems and their teleological devices is the fact that *any intentional action (intention to do) when put into execution must be implemented at a lower layer of not really intentional sub-acts*, merely automatically adjusted and just retrieved from our action-repertoire.

The schema we proposed for such integration/implementation is the following one:

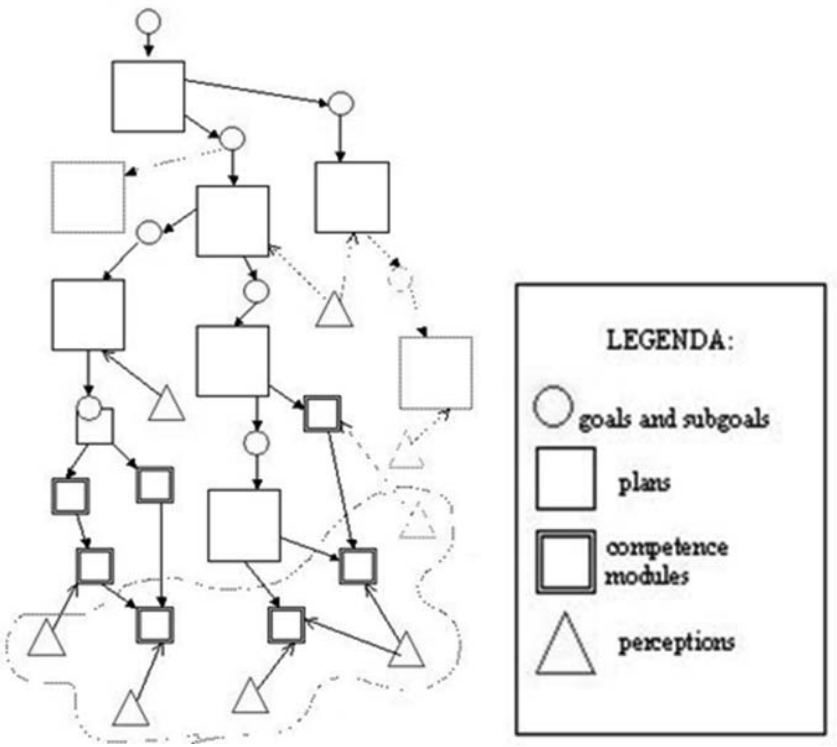

Fig. 3 Functional continuum between Intentional Goals and automatic Classifiers

There is a functional continuum: The top part is more similar to the BDI (Beliefs Desires Intentions) model (Rao and Georgeff 1995; Bratman 1987). The lower part is more similar to Behavior Networks (Maes 1989) and uses anticipatory classifiers (Pezzulo et al. 2008; Pezzulo & Castelfranchi 2009). Executive Intention ("Intentions in action") are/must be *implemented* in lower structures (production rules, reflexes,

classifiers), which, when specified, are represented in sensory-motor images/schemes.

For example, the intention to open the door is executed by a lot of micro-actions (bend our fingers, pull, move our feet to pass) which are not "intentional" but finalistic schemas. When I do intentionally take a walk I do not intentionally bend my feet. (See Dunja Jutronić's paper in this volume.)

5. *Considerations on "intentions" in homage to Kathy Wilkes*

"Intentions" only in language using organism? To discuss this thesis we have first to make clear what kind of goal are "Intentions" in our model and where they derive from.

5.1. *What "Intentions" are: a kind of Goal*

"Goals" and "Motives" do not mean "Desires." It is not synonym of "goal" like in Bratman's BDI model (Bratman 1987). Desires are just one *kind* of goal. Desires are endogenous (and usually pleasant) and with "norms" we have just to cut some possible course of action by *making some desire of the subject practically impossible or non-convenient*. Intentions do not derive just from "desires" but also from other kind of goals. They can derive from norms, prescriptions, *duties*; but *"duties" are not "desires";* they are *goals from a different source*, with a different origin: they come from outside (*exogenous*),[8] they are imported, "adopted," they are "prescriptions" and "imperatives" from another agent.

Not all goals have to be "(actively) pursued," like for "intentions." A goal is not a goal only if/when pursued. Some of them (like having a sunny day) are not within our power: to realize them is not up to us, but depends on other "agents" or external forces, thus we cannot really "pursue" them. Other goals are just partially up to us; we have to do something but then the final result depends on the others, or on luck. Thus, we may have actively pursued goals (goals pursued through our active actions), but also merely passive goals; and the latter can be of two very different kinds:

- goals we have just to wait for, to hope for their attainment; which do not depend at all on us: we cannot do anything (else).
- goals whose realization depend on us and on our "doing nothing," that is abstaining from possible interference. We would have the power to block that event/result, and we decide to do nothing in order to let it happen (inaction, "passive action").

Furthermore, because not all goals are directed towards *approaching* a desirable outcome goals can also be directed towards *avoiding* an

[8] However, see later about the internalization of the "authority" and internal moral imperatives.

undesirable outcome (Elliot 2006). Avoidance and approach represent two mental frames, two different psychological dispositions and mind settings (see Higgins' avoidance and approach "regulatory *focus*" in his 1997).

Not all our goals are "felt" because not all of them are represented and defined in a sensory-motor format (see below).[9] The two most important kinds of felt goals are *desires* and *needs*.

Intentions are those goals that *actually drive our voluntary actions or are ready/prepared to drive them*. They are not another "primitive" (like in BDI model), a different mental object with respect to goals. They are just a kind of goal, the final stage of a successful goal-processing with very specific and relevant properties (see Castelfranchi and Paglieri 2007).

In a nutshell, in our model, an *intention* is a goal that:

1) has been activated and processed;
2) has been evaluated as not impossible, and not already realized or self-realizing (achieved by another agent), and thus *up to us*: we have to act in order to achieve it;[10]
3) has been chosen against other possible active and conflicting goals, and we have "decided" to pursue it;
4) is consistent with other intentions of ours; a simple goal can be contradictory, inconsistent with other goals, but, once it is chosen, it becomes an intention and has to be coherent with the other intentions;[11]
5) implies to the agent's belief that she knows (or will/can know) how to achieve it, that she is able to perform the needed actions, and that there are or will be the needed conditions for the intention's realization; at least the agent believes that she will be able and in condition to "try";
6) being "chosen" implies a "commitment" with ourselves, a mortgage on our future decisions; intentions have priority over new possible competing goals, and are more persistent than the latter (Bratman 1987);
7) is "planned"; we allocate/reserve some resources (means, time, etc.) for it; and we have formulated or decided to formulate a plan con-

[9] We mean that, for example, we cannot say "I feel the intention of…" simply because the sensory-motor format of the represented anticipatory state is not specified in the very notion of "intention." "Intention" is a more "abstract" representation, and kind of goal, with a non-specified codification. Looking at a goal as an "intention," we abstract away from its possible sensory components.

[10] An intention is always the intention to "do something" (including inactions). We cannot really have intentions about the actions of other autonomous agents. When we say something like "I have the intention that John goes to Naples" what we actually mean is "I have the intention *to bring it about that* John goes to Naples."

[11] Decision-making serves precisely the function of selecting those goals that are feasible and coherent with each other, and allocating resources and planning one's actual behavior.

sisting of the actions to be performed in order to achieve it. An intention is essentially a two-layered structure:
(a) the "intention that," the *aim,* that is, the original processed goal (for example, to be in Naples tomorrow);
(b) the "intention to do," the sub-goals, the planned executive actions (to take the train, buy the tickets, go to the station, etc.). There is no "intention" without (more or less) specified actions to be performed, and there is no intention without a motivating outcome of such action(s).

8) thus, an intention is the final product of a successful goal-processing that leads to a goal-driven behavior.

After a decision to act, an intention is already there even if the concrete actions are not fully specified or are not yet being executed, because some condition for its execution is not currently present. Intentions can be found in two final and pre-final stages:

(a) *Intention "in action,"* that is, guiding the executive "intentional" action;
(b) *Intention "in agenda"* ("future directed" intentions, those more central in the theories cited of Bratman), that is, already planned and waiting for some lacking condition for their execution: time, money, skills, etc. For example, I may have the intention to go to Capri next Easter (the implementation of my "desire" of spending Easter in Capri), but now is February, and I am not going to Capri or doing anything for that. I have just decided to do so at the right moment; it is already in my "agenda" ("things that I have to do") and binds my resources and future decisions.[12]

5. 2. *"Intentions" only in language using organism?*

A very crucial thesis of Kathy Wilkes is her conclusion that "goal-representation can only be ascribed to language using organism" also due to her caring/stressing the distinction between "intentionality" and "intensionality." This is a crucial distinction. However, I disagree about that conclusion/thesis, which refers to "intentions." My first point is that "mental representations" are not only in linguistic format and based on language. We also have another kind of "mental" representation and mental working, i.e. *mental images and to imagine.*[13]

Also this kind of representations are really semiotic, have their "semantics" (content/object/aboutness). Not only Knowledge (epistemic representations) but also Goals (motivational representations) can be mentally represented *in sensorymotor format, as mental images*. Paradoxically, the example used by Miller, Galanter, and Pribram in their

[12] I would also say that an "intention" is "conscious," we are aware of our intentions and we "deliberate" about them; however, the problem of unconscious goal-driven behavior is open and quite complex (see Bargh et al., 2001).

[13] We know that for Piaget the first level of "intelligence" and thinking is precisely "sensorymotor thinking."

famous book was a nail driven into the wall, where the Goal was a mental Image compared with the perceived one.

However, there may be a possible convergent hypothesis with Kathy Wilkes's challenging claim. As we said, "Intention" in the strict sense belongs to the domain of System 2 and is the result of the "deliberative" processing. It is the result of reasoning, preference and choice *based on* "*arguments,*" *reasons*, that is beliefs supporting one goal or the other; *it can be explained, discussed*. I can even discuss and argue with my own self; but this doesn't imply that the intended goal/objective itself is formulated in linguistic format.

However, the creation of the "intention" also *entails beliefs about the Agent itself*: my skills, know how: "Am I able to; Do I know what/how to do?". And what about my own mind? Perhaps this self-representation should be expanded: it might entail some *meta-cognitive* representation: not only Beliefs about my mind but meta-Goals.

Since an Intention is a goal about my own agency, my performing an action, it might imply *a goal not only about my doing something but about my having the goal*: a goal about my mind and my commitment. But how can this be formulated, represented a goal about my having a goal, my mind?

5.3. *From "Intention" to "Volition"?*

While the goal of doing/performing a given action can be still formulated in sensory-motor format representation, such goal about my goal/mind reasonably would need a linguistic/communicative representation (a reflexive sociality). This is for me the possible point of convergence/agreement with Kathy Wilkes's thesis.

Not the goal/object of the intention and intentional action is necessary linguistic, it can be merely sensory-motor image (like empting my glass; turning off the stove), but its meta-cognitive, reflexive component is linguistic. *A goal about my mind, my having a goal*, must be represented in an abstract, propositional, conceptual form.

However, I would say that this is no longer just "intention" but it is a "will" and a voluntary action controlled by will; a stronger form of intentionality where I'm *socially influencing my-self* (and language and (self-)mind reading are for that). In fact, the so- called "strength of the will" is my influencing power over my-self: to impose my own self to do something and to be committed, and to control myself.

References

Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A., Barndollar, K. and Trotschel, R. 2001. "The automated will: Unconscious activation and pursuit of behavioral goals." *Journal of Personality and Social Psychology* 81: 1004–27.

Bargh, J. A. and Ferguson, M. J. 2000. "Beyond behaviorism: The automaticity of higher mental processes." *Psychological Bulletin* 126: 925–45.

Bratman, M. 1987. *Intention, Plans, and Practical Reason*. Cambridge: Harvard University Press.

Castelfranchi, C. 1998. "Il nevrotico cripto-utilitarista: contro l'ideologia del 'vantaggio secondario'". *Sistemi intelligenti* 10 (2): 307–314.

Castelfranchi, C. 1999. "La fallacia dello psicologo. Per una teoria degli atti finalistici non intenzionali." *Sistemi Intelligenti* 11 (3): 435–68.

Castelfranchi, C. 2001. "The theory of social functions. Challenges for multi-agent-based social simulation and multi-agent learning." *Journal of Cognitive Systems Research* 2: 5–38.

Castelfranchi, C. 2012a. "Goals, the true center of cognition." In F. Paglieri, L. Tummolini, R. Falcone, M. Miceli (eds.). *The Goals of Cognition*. London: College Publications.

Castelfranchi, C. 2012b. "'My mind'. Reflexive sociality and its cognitive tools." In F. Paglieri (ed.) *Consciousness in Interaction: The role of the natural and social context in shaping consciousness*. Amsterdam: John Benjamins, 125–150.

Castelfranchi, C. 2017. "Goal 'Value': Not just 'Dual' but 'Hybrid.'" In T. Everitt, B. Goertzel, A. Potapov (eds.). *Artificial General Intelligence: 10th International Conference*. Melbourne, 45–54.

Castelfranchi, C. and Paglieri F. 2007. "The role of beliefs in goal dynamics: prolegomena to a constructive theory of intentions." *Synthese* 155 (2): 237–263.

Conte, R. and Castelfranchi, C. 1995. *Cognitive and Social Action*. London: UCL Press.

Elliot, A. 2006. "The hierarchical model of approach-avoidance motivation." *Motivation and Emotion* 30 (2): 111–116.

Higgins, E. T. 1997. "Beyond pleasure and pain." *American Psychologist* 52: 1280–1300.

Jutronić, D. 2022. "Intentions and their role in (the explanation) of language change." *Croatian Journal of Philosophy* 22 (3): 327–350

Loewenstein, G. and O'Donoghue, T. 2004. "Animal Spirits: Affective and Deliberative Processes in Economic Behavior." *Microeconomic Theory eJournal*. https:// cpb-us-e1.wpmucdn.com/blogs.cornell.edu/dist/b/5495/files/2015/10/will5_05-227fjlg.pdf

Miller, G.A., Eugene Galanter, E. and K. H. Pribram. 1960. *Plans and the Structure of Behavior*. New York: Henry Holt and co.

Pezzulo, G., Butz, M., Castelfranchi, C. 2008. "The Anticipatory Approach: Definitions and Taxonomies." In G. Pezzulo, M. V. Butz, C. Castelfranchi and R. Falcone (eds.). *The Challenge of Anticipation: A Unifying Framework for the Analysis and Design of Artificial Cognitive Systems*. Cham: Springer 2008, 23–43.

Pezzulo, G. and Castelfranchi, C. 2009. "Thinking as the Control of Imagination: a Conceptual Framework for Goal-Directed Systems." *Psychological Research* 73: 559–577.

Viale R. 2018. "The normative and descriptive weaknesses of behavioral economics-informed nudge: depowered paternalism and unjustifed libertarianism." *Mind and Society* 17: 53–69.

Viale R. 2019. "Architecture of the mind and libertarian paternalism: is the reversibility of system 1 nudges likely to happen?" *Mind and Society* 18: 143–166.

Croatian Journal of Philosophy
Vol. XXII, No. 66, 2022

Book Review

Jessica Brown, Fallibilism: Evidence and Knowledge. *New York: Oxford University Press, 2018, 197 pp.*

If I were lucky enough to enjoy the experience of sitting in a pub with a couple of friends on a Friday night, I would certainly not complain. But would I be justified to claim that I *know* that I am in the pub with them? I just might be dreaming or hallucinating this pleasant event. Conversely, if I were to sit at my desk having proved that 7+5=12 by relying on Peano's axioms, would I be able to say that I *know* this to hold for my system? Surely there is a substantial difference between the two situations. It appears to me that although I might be dreaming that I proved this simple mathematical claim, it is not possible that I am *not* in the state of knowing that it holds. Of course, the grand majority of my beliefs are more similar to the former situation than the latter. Most of my beliefs are about my experiences, not formal mathematical proofs. And even though it appears I am much more inclined to say that 7+5=12 holds than that I am, in fact, in the pub with my friends, I would want to say that I know both these things.

In her book, *Fallibilism: Evidence and Knowledge*, the author Jessica Brown tackles this issue from a fresh perspective, as she recognizes that the problem of the explanatory gap between evidence and knowledge has been central to the 20th and 21st-century epistemology. As many philosophers had taken a stand in saying that one's evidence for p can rarely conclusively establish that p, the concept of knowledge was shown to be quite troublesome. How am I to say that I *know* that p without possessing conclusive evidence that p? Or, in Brown's formulation, how am I to say that I *know* that p if p might be false? Three positions are widely advocated in their respective attempts to answer this question: (1) fallibilism, exemplified by the claim that one can know that p while retaining the possibility of p being false, i.e., evidence not guaranteeing that p, (2) infallibilism, exemplified by the claim that one can know p only if their evidence conclusively points to p, and finally (3) skepticism, claiming that the gap between evidence and knowledge is unbridgeable, and hence that one can, in fact, know very few things, if any.

As one can make an educated guess from the book's title, Jessica Brown has opted for the fallibilist account of knowledge. Throughout the course of 8 chapters, she examines the most persuasive accounts of fallibilism and infallibilism in their respective attempts to navigate the epistemic battlefield, managing somehow not to fall into the skeptic's trench of unknow-

ability. Brown's representation of opposing theories is very bona fide; her arguments are clear and do not seem to obfuscate the matter. The book's preface offers a simple yet informative guide for the reader, presuming only the basic knowledge of concepts in contemporary epistemology. The organization of chapters is also well-thought-out and easy to follow, as each chapter tackles a discrete point in the discussion. The transition between the chapters is also often seamless, making reading the book quite pleasurable.

Before we get into the overview of the chapters in the book, a couple of points of terminological clarification ought to be made. The author uses the term *shiftiness* to describe the original conception of knowledge in the infallibilist theory, proposed by Lewis in 1996. Although in itself quite problematic, this account gave a new rise to infallibilist theories at the end of the century, which have since become quite dominant. Lewis's shifty knowledge, as Brown describes it, is closely bound to the theory of epistemic contextualism, which claims that the attribution of knowledge depends, at least to some degree, to something in the context of the person who attributes knowledge to the subject. For this reason, epistemic contextualism is often referred to as attributor contextualism. This is basically why Brown uses the term shifty conception of knowledge, as non-context dependent theories of knowledge are, in essence, invariantist. In other words, invariantism promotes universal theory of knowledge attribution. The other concept that probably needs some clarification is a *generous conception of evidence*. Now, what exactly does generosity have to do with evidence? It stands to reason that to bridge the obvious gap between evidence and knowledge, one might try either weaken the concept of knowledge, as was the case with Lewis's contextualism, or opt for reframing the concept of evidence. If the conception of evidence is rendered inclusive enough, the gap will be closed. This kind of manoeuvre stretches the conception of evidence from covering only the claims about our experieces to claims about the external world as well. If one has no problem attributing our claims' content from the external world, bridging the gap might be quite an unproblematic task. But more on this later on.

It would appear useful to actually get to know Brown's main opposition, the authors who will attempt to defend infallibilist theories regarding evidence and knowledge. Even though they can be viewed as proponents of the same theoretical position, their respective views on how to attain the infallibilists' goal of bridging the aforementioned gap are, in fact, very different. As I have already briefly touched upon Lewis's contextualist attempt to construct a shifty knowledge-based theory, it would be best to turn our attention to the other couple of authors that Brown cites as representative of their respective approaches. The first of them is John McDowell, whose disjunctivist epistemology opens the door for the infallibilist position. As Brown eloquently put it: "Disjunctivists about experience hold that the state of its looking to one as if p may be constituted either by one's seeing that p or it's merely appearing to one as if p" (3). McDowell continues on this line of argumentation by claiming that in optimal conditions when one is in a state of experiencing something, "it is a matter of the fact itself being disclosed to the experiencer." Such a position obviously allows McDowell to claim a non-shifty non-sceptical infallibilism, however Brown sees his

conception of evidence as being much too inclusive, as will be evident in her criticism. The other author discussed by Brown in the book who attempts to construct a non-shifty non-skeptical account of infallibilism is no other than Timothy Williamson. His knowledge-first program considerably impacted the contemporary discourse of epistemology by giving knowledge explanatory priority when addressing the process of epistemic justification. And although put in the same basket of infallibilism, his approach radically differs from one taken by McDowell. Williamson claims that the subject's knowledge, in fact, *is* subject's evidence. If that holds, the consequence is that when one is in a state of knowing that p, then p is his evidence that p. This entailment, as Brown says, makes p's probability 1. In other words, possessing knowledge that p guarantees p, making his position unambiguously infallibilist.

In this book, Jessica Brown chooses to attack both accounts of non-shifty non-skeptical infallibilism by claiming that their liberal approach to the concepts of evidence and knowledge leads to undesirable philosophical implications. She also recognizes that the objections made to the fallibilist theories also hold for the infallibilist ones. Her considerations finally push forward the idea that if both groups of theories, fallibilist and infallibilist, generate virtually the same philosophical problems, one should opt for fallibilism as it at least doesn't stretch the concepts of evidence and evidential support unnecessarily.

Now that we have settled the basics of the discussion, let us turn to a short overview of the chapters in the book. The first chapter elaborates on the positions of fallibilism and infallibilism, with Brown selecting the most persuasive accounts of both worlds, at least in her own view. She examines the motivations behind infallibilism and claims that the main one is the unintuitive view of the fallibilists that one can *know* p while maintaining that p might not be true. In short, the first chapter is mainly expositional, setting up the stage for arguments of both sides.

The second chapter deals with the account of infallibilism that she chose to address, claiming that the externalist commitments made by its proponents in the context of evidential support are largely untenable. The three commitments she recognizes as philosophically and intuitively problematic are: (1) factivity, the commitment to p being evidence only if it is true, (2) sufficiency of knowledge for evidence, the commitment to the claim that if S knows that p, then p is a part of S's evidence, and finally (3) sufficiency of knowledge for self-support, the commitment to the claim that if S knows p, then p constitutes, at least in part, evidence *for* p.

As Brown introduced these three infallibilist commitments in the second chapter, she decided to focus on the commitment of sufficiency of knowledge for self-support in the third chapter. She specifically challenges this commitment by claiming the theorists who accept it has to answer the question of why it usually appears infelicitous to have p as evidence for itself. She attempts to see if this commitment is defensible by accepting one of the probabilistic accounts of evidential support but ultimately deems them quite controversial.

The fourth chapter constitutes her final case against accepting infallibilism by putting forward an argument which questions factive concep-

tion of evidence, viz. knowledge constituting evidence only if it is true. She supports this by appealing to the thought experiment of a subject and its counterpart BIV, who share some experience which adequately represents the state of affairs in the world for the subject, but not for BIV. For example, let us imagine that both the subject and BIV have the experience of eating dinner; however, only the subject's experience, in fact, corresponds to what is going on. By accepting the commitment of factivity, one ought to say that only the subject is justified in his belief, being right about his belief. She notes that the infallibilists attempt to defend this commitment by claiming the strawman fallacy in opposition's argument; they state that the opponents criticised equal blamelessness in accepting a belief instead of equal justification. She argues that the defense is unsuccessful in its endeavor since it fails to recognize that "on the knowledge view of justification, justification cannot play key roles traditionally played by justification, including providing a graded and propositional notion of justification" (22).

In the fifth chapter of the book, Brown settles accounts with the principle of epistemic closure, which is often seen as one of the more appealing reasons for accepting infallibilism. She rightly argues that if the principle of closure fails due to some external reason, it becomes irrelevant which theory, fallibilist or infallibilist, is better calibrated for it. She attempts to show that the closure principle fails due to epistemic defeat, which means that the introduction of new information can cause the existing beliefs to lose ground in their respective justifications. This chapter probably offers more contribution to the discussion than any other in the book.

The sixth chapter capitalizes on an epistemic defeat that Brown advocates, with a focus on the undermining defeat. This type of epistemic defeat consists of the subject being provided new information that renders their justification process of a belief invalid, but does not support the opposite claim either. She considers so-called level-splitting views that are based on higher-order evidence which should inhibit the justification of subject's beliefs, but ultimately concludes that they result in untenable accounts of theoretical and practical reasoning, making them philosophically problematic.

The seventh chapter constitutes Brown's defense of fallibilism when faced with its difficulty handling practical reasoning and concessive knowledge attribution. This problem for fallibilism is often used as a reason for accepting infallibilism; however, she again makes her case by showing that infallibilism faces the same issues and argues that both positions have a wide array of options and adequate instruments for dealing with them.

Finally, in the last chapter of her book, she provides a comprehensive summary of reasons for accepting fallibilism, despite criticism often thought to be detrimental to the theory. She argues that both fallibilists and infallibilists have much room for maneuver in defending their respective theories.

ANTE DEBELJUH
University of Rijeka, Rijeka, Croatia

Table of Contents of Vol. XXI

Book Reviews

Croatian Journal of Philosophy is published three times a year.
It publishes original scientific papers in the field of philosophy.

Croatian Journal of Philosophy is indexed in
The Philosopher's Index, PhilPapers, Scopus, ERIH PLUS
and in *Arts & Humanities Citation Index* (*Web of Science*).

Payment may be made by bank transfer

SWIFT PBZGHR2X
IBAN HR4723400091100096268

Croatian Journal of Philosophy is published with the support
of the Ministry of Science and Education
of the Republic of Croatia.

Instructions for Contributors

All submissions should be sent to the e-mail: cjp@ifzg.hr. Submissions must be in English and formatted to be double-spaced with suitably wide margins, an A4 page size, and automatic page numbering.

Articles are normally no longer than 8,000 words of main text including bibliography. The Journal will consider longer papers, but, once these limits are exceeded, authors should bear in mind the editorial policy that the acceptance bar raises with increasing length.

Manuscripts should be compiled in the following order: cover page; title; abstract (not exceeding 200 words); keywords (3 to 6); main text; appendices (as appropriate); references.

All the authors of a paper should include their full names, affiliations, postal addresses, telephone and fax numbers and email addresses on the cover page of the manuscript. If a paper is co-written, one author should be identified as the Corresponding Author. The cover page must be submitted as a separate document. All submitted manuscripts must be prepared for blind review, with revealing acknowledgements and self-identifying references removed.

Sources are cited in the text by the author's last name, the publication date of the work cited, and a page number if needed, e.g. (Barber 2007: 324). Full details appear in the reference list in which the year of publication appears immediately after the author's name:

Barber, A. 2007. "Linguistic Structure and the Brain." *Croatian Journal of Philosophy* 21 (7): 317–341.

Williamson, T. 2013. *Identity and Discrimination*. Oxford: Wiley-Blackwell.

The publication of a manuscript in the Croatian Journal of Philosophy is expected to follow standards of ethical behavior for all parties involved in the publishing process: authors, editors, and reviewers. The journal follows the principles of the Committee on Publication Ethics (https://publicationethics.org/resources/flowcharts).

ISSN 1333-1108
9 771333 110001