New College of the Humanities

British Wittgenstein Society Conference

Contributed Abstracts

Wittgenstein and AI

New College of the Humanities

Wittgenstein and the Formalization of Complexity

Serena Dell'Unto (Pontifical Gregorian University, Rome)

29 July 14:00

Wittgenstein's mature philosophy taught us to assume a perspective that understands language as a complex of communicative interweaving. Unlike the representational conception of the language- system, which refers to the Tractatus, the Philosophical Investigations show the complexity, vagueness and absence of fixed boundaries in a language functional to the needs of life. Such language, significant in all its forms, exhibits a certain normativity1, which however does not allow for the anticipation of every single linguistic move of the speakers. This normativity keeps the language in a dynamic equilibrium, equally distant from both extreme fixism, resulting from the imposition of absolute rules, and extreme chaos, which would instead be determined by the absence of rules. It is about understanding whether AI and, in particular, computational linguistic systems can model the complexity of a language that relies on this type of normativity. Such systems need to adapt to multiple linguistic contexts, to manage unexpected events and strategically fit into the complex intertwining of life forms. In the context of computational linguistics, among the bottom-up models, cross-situational word learning2 seems to be the most adaptive one. It is based on a probabilistic mechanism that allows the system to progressively refine its knowledge after repeated exposure to contexts or modify it after different stimuli. It is a model that manages referential ambiguity3 and any potential distractor. It has a theoretical framework very close to Wittgenstein's intuitions and it is highly functional to reproduce computational systems capable of producing and understanding language. Despite the enormous potential, this model has a few limitations, such as the main focus on the denotative function of language, not considering that, as Wittgenstein teaches, this function is one of the many functions that language presents and not even the most important one. If the cross-situational learning model, in addition to the denotative function, took also into account the links between words and their regular distribution in language4, it would allow the system to achieve higher and more effective performances. The challenge that computational linguistics propose is to create strategies to manage a "fine-grained" reality, seeking a formalization that does not reduce the complexity of reality. With the warning «Don't think, watch!»5, Wittgenstein removed the distorting lenses that have traditionally circumscribed reality in predefined schemes and showed the richness of a living and intrinsically dynamic language. His method can be a guide, providing regulating ideas that direct AI towards adaptive, complex, and intelligently efficient systems.

1 For the link between meaning and normativity see S. Kripke, Wittgenstein on Rules and Private Language.
2 A. Fazly, «A Probabilistic Computational Model of Cross-Situational Word Learning».
3 On referential uncertainty see W. Quine, Word and Object.
4 A. Lenci, «Distributional Models of Word Meaning».
5 L. Wittgenstein, Philosophical Investigations, §66.

New College of the Humanities

Toward a Wittgensteinian Cognitivism

David Lindeman (Georgetown University)

29 July 14:40

With its picture theory of meaning, Wittgenstein's Tractatus provides an early articulation of truth-conditional semantics. Partly under the influence of Noam Chomsky, in turn partly under the influence of later Wittgenstein, Paul Pietroski argues that we ought to abandon truth-conditional semantics. On Pietroski's alternative, sentence meanings are not truth-conditions. Instead, sentence meanings are instructions for composing concepts in a (Fodorian) language of thought. At first blush, this cognitivist approach to semantics looks like a significant departure from the Tractarian line of thought. But Pietroski allows that (at least certain stretches of) this language of thought (Mentalese) comprising the concepts composed in executing the instructions that meanings are may have a truth-conditional semantics; and implicit in the Tractatus, I argue, are the outlines of a cognitivist account of propositionally-articulated thought agreeing in all essential respects with Pietroski's full account—provided that a Mentalese sentence composed by executing the meaning of a natural language sentence and that natural language sentence share a logical form. Drawing on Chomsky's competence/performance distinction, and recognizing that what is provided in the Tractatus is not a performance model but the outlines of a competence model, we arrive at a form of psychologism, but not one that gets us "entangled in unessential psychological investigations".

New College of the Humanities

Meaning as Use and Distributional Semantics

Jumbly Grindrod (University of Reading)

29 July 15:20

Distributional semantics is a widely-used approach for capturing linguistic meaning within artificial intelligence. It has proven to be crucial to the state-of-the-art across a range of natural language processing tasks, and it lies at the core of widely-discussed language models like Google's BERT and OpenAI's GPT-3. The approach is often initially justified via appeal to the distributional hypothesis: that terms similar in meaning will be similar in their distribution across a corpus, and likewise that terms dissimilar in meaning will be dissimilar in their distribution. The metasemantic view of later Wittgenstein that "the meaning of a word is its use in the language" (Wittgenstein, 1953, 43) is often appealed to as a theoretical underpinning for the distributional approach (Firth, 1957; Lenci, 2008, 2018; Sahlgren, 2008; Erk, 2012; Westera and Boleda, 2019). However, this idea is rarely developed beyond the slogan that meaning is to be understood in terms of its use. In this talk, I will explore whether we can understand the distributional approach via the use-based metasemantic view of later Wittgenstein. In particular, I will consider:

- The difference between use as Wittgenstein conceived it vs use as represented by a corpus.
- The extent to which a corpus-based approach removes worries about delineating language games
- The possibility that distributional approaches provide a new way of understanding linguistic meaning without appeal to necessary and sufficient conditions
- Finally, whether the distributional approach provides a way of rooting meaning in use without appeal to a cognitive turn i.e. without viewing a theory of meaning as a cognitive theory of speaker understanding.

New College of the Humanities

AI and the Cluster Account of Art

Alice Helliwell (New College of the Humanities)

29 July 14:00

The first question asked by those presented with artworks made by AI is frequently “but is it really art?”. This paper offers a way of answering this question, through appeal to the neo-Wittgensteinian cluster account of art. A Wittgensteinian approach to defining art was first proposed by Morris Weitz (1956). Weitz argued that we should reject attempts to offer sufficient and necessary conditions for art. Weitz instead proposes that art, like games, is a concept better elucidated by resemblances: ‘If we actually look and see what it is that we call “art”, we will also find no common properties—only strands of similarities’ (Weitz 1956, 31). This approach to art was broadly rejected in aesthetics, in part due to the difficulty of finding satisfactory paradigm cases (Gaut 2005). This family resemblance approach to art was resurrected by Berys Gaut in 2000 as a cluster account of art. Since it was defended by Gaut (2005) the cluster account of art has maintained its status as a plausible alternative to definitionalist approaches. In this paper, I will examine how compatible the cluster account of art is with AI artworks, in comparison to prior definitionalist accounts. I will begin by presenting examples of generative artworks, and the AI systems which produce them. I will then contrast the approaches of definitionalist accounts of art with the cluster account. I will finally demonstrate the flexibility of a family resemblance approach in allowing AI artworks to be considered art, and raise a potential issue for AI art under Gaut’s formulation. I aim to show: 1) that a cluster account of art can help us to understand how AI artworks may fit into the category of ‘art’, and 2) that a cluster account can also clearly show us the common features of art which AI artworks cannot (yet) reach, without preventing it from being ‘art’ at all. Under Gaut’s account, we may also find that some AI works will count as fringe cases, giving us cause to consider them at the very least as art-like. The cluster account of art offers us a useful approach to analysing and potentially accommodating not only new genres or mediums of art, but new forms of artist.

New College of the Humanities

“The Forms of Artificially Intelligent Life”: Brandom, Chomsky, and Wittgenstein on the Possibility of Strong-AI

Laith Abdel (Independent Researcher)

29 July 14:40

The impact of artificial intelligence on our day-to-day lives cannot be understated. From Alexa to GPT-3, Computer Science contributed remarkable implementations of AI which has changed our lives forever. Natural Language Processing, Machine Learning, and Automated Reasoning have garnered special attention in recent years due to their potential in the development of the world’s first Strong-AI. All models of AI currently rely on a statistical approach to the problem of automating the interpretation of large sets of data not previously processed by a program. I call the current paradigm “Statistical AI”. This method is based on a conception of human reasoning similar to that of Behaviorism. For instance, Backpropagation is a popular model used in Neural Network Processing and is a proper statistical model of conditioning an agent to interpret an action based on reward and punishment. Even Unsupervised Learning methods rely upon similar statistical approaches. Is Statistical AI the key to finally creating a Strong-AI, or misled by its adoption of a computational version of Behaviorism? This paper examines these questions through the lenses of Robert Brandom, Noam Chomsky, and Ludwig Wittgenstein. First, I apply Chomsky’s Universal Grammar and Semantic Nihilism as an attack on Statistical-AI. Universal Grammar (UG) is a biological, innate set of constraints that determine the structure of syntactic relations in human grammars. UG only needs “external systems to provide minimal design specifications.” (Chomsky 2000) There is no causal relation between semantics and the innate syntactic structures constrained by UG. By contrast, Statistical AI is a system that is entirely dependent on external data to generate a proper algorithmic function to correctly interpret future data. Statistical AI cannot develop a Strong-AI without first successfully modeling UG. Then, I interpret Brandom’s Analytic Pragmatism and Pragmatic AI as agnostic towards Statistical AI. Statistical AI could not be a Strong-AI if it cannot participate in a discursive community to validate its assertions. For Brandom, this issue is “the biggest technological challenge to getting computers either to be able to participate as full-fledged members of our discursive communities or, us, programming them to be able to form their own communities which would confer content.” (Brandom 2019). In response to Chomsky and Brandom, I refer to Ludwig Wittgenstein’s work on Internalism and Externalism. In my reading of Wittgenstein, I find his views sympathetic to particular claims of Chomsky and Brandom while rejecting their respective Internalist and Externalist foundations. The necessary conditions that Strong-AI must satisfy are based upon our concept of human cognition, to which neither Science nor Philosophy have a clear objective view. It is a “groundless ground” that current Computer Scientists and Philosophers alike remain fundamentally anthropocentric in their endeavors to mimic or replicate the human mind through models. The paper then ends with a discussion on how Computer Science can learn from the insights of the three views and how AI could move forward or regress in its endeavors towards bringing Strong-AI to life, on the grounds that it must resemble human beings and our forms of life.

New College of the Humanities

The Metonymical Trap

Éloïse Boisseau (University of Aix-Marseille)

29 July 16:10

Actions performed by machines are philosophically puzzling. If there seems at first to be no difficulty in ascribing some powers to machines (we say that machines wash clothes, drill holes, compute numbers, etc.), the question remains whether it makes sense to say of the actualization of these powers (for instance the washing of the clothes, the drilling, or the computing) that it is the result of machines taking action (Hacker, 2007), or more generally if it is a case of machines acting in the full sense of the word. Following Wittgenstein's (1958) notorious distinction between two different types of questions regarding machines, viz. logical ones ('Is it possible for a machine to think?') and empirical ones ('Can a machine liquefy gas?'), I would like to try and shed light on the question of the qualification of the doings of evermore complex machines by investigating the logical descriptions of their actions and powers. Centrally, the idea I would like to put forward is the following: the more complex machines become, the more we seem to forget that they are precisely nothing more than that (machines), and the more we feel that they are worthy of qualifications we ordinary reserve to animate agents (especially human beings). This tendency towards the ascription of cognitive abilities and agency to machines is not benign and might be explained in part due to what I shall call a 'metonymical trap'. A metonymy is a figure of speech in which something is being referred to by using a word that is closely related to it. The metonymical trap is close to, but comparatively more general than, the homunculus fallacy (Kenny, 2008) or the mereological error (Bennett & Hacker, 2003). I will argue that we indeed fall into a metonymical trap when we attribute to a machine the actions performed by the human being who is actually using the machine. One should instead, expanding Proudfoot (2013), distinguish between an acting-machine and a machine that acts and ascertain that, while there are certainly acting-machines, it does not make sense to say that there are machines that act. What I want to establish is that although there indeed exists a common form of description of one action performed both by a human being and through the use of a machine (a description using mechanical terms), this description only consists in one form of description of the action and does not exhaust its full scope of characterization, especially because it does not consider the action in its intentional dimension (Anscombe, 1979; Kenny, 2003).

New College of the Humanities

Can Machines Act Ethically?

Luca Alberto Rappuoli (University of St Andrews)

29 July 16:50

The point of engaging in ethical reflection lies in the simple yet intractable question 'How should we live?'. According to McDowell (1979), this question cannot be approached by identifying a set of rules of conduct to govern our actions. Any such endeavour is in fact condemned to failure for — he argues — our moral outlook is simply not susceptible of codification in a finite set of rules. The impossibility of codifying our moral outlook in a finite number of principles might strike us as a plausible thesis. Yet — upon reflection — we can detect a notable tension between this uncodifiability and our intuitions about rationality. Indeed, it appears that acting rationally (on the basis of reasons) can only be explained in terms of following formulable rules. Now — the thought continues — our moral outlook constitutes a specific reason for acting as we do and, as such, must therefore be explainable in rule-following terms. Nonetheless — McDowell argues[1] — Wittgenstein has compellingly shown that a rule-following conception of rationality is nothing but a deep-rooted prejudice.[2] The central idea is that when we posit a rule, there is no guarantee that our audience will act in accordance with the same psychological mechanism we intended to posit. At any given time, the evidence at our disposal is in fact compatible with a future change in the behaviour of the audience. Hence, the mere postulation of a rule cannot guarantee the uniformity of behaviour required by an exercise of rationality. In McDowell's eyes, Wittgenstein's considerations clear the way for accepting the idea that our moral outlook is essentially irreducible to a set of rules.[3] That is, for every finite set of moral rules, there will always be a situation in which a mechanical application of the rules will result in an action that a perfectly ethical agent would not have performed in that same situation. In this paper, I shall argue that embracing this thesis would lead us to accept the incredibly committing idea that there can never be an AI that could be properly characterised as 'ethical'. Indeed, the nature of AI is entirely underpinned by rule-following processes. More precisely, due to its algorithmic nature,[4] AI is constitutively incapable of acting in a way that resists a description in terms of acting in the light of a rule. It follows that, if the moral outlook of an ethical agent is constitutively uncodifiable, there is no possible set of rules on which the AI can rely to render its actions consistently similar to that of the ethical agent. Wittgenstein's considerations, however, appear insufficient to reach the incredibly committing conclusion that there will never be an ethical AI. This tension, I suggest, should be resolved by employing Wittgenstein's remarks on rule-following as an argument for the constitutive irreducibility of our moral outlook to propositional knowledge (epistemic opacity) — rather than as an argument for its uncodifiability.

1 Cf. (McDowell 1979: 337-338).
2 See (Wittgenstein 1953: §185).
3 Cf. (McDowell 1979: 339).
4 Cf. (Russell and Norvig 1994: 1024-1026).

New College of the Humanities

Autonomy: a Family Resemblance Concept; Evidence from Robotics

Emily Collins (Northeastern University)

29 July 17:30

What is autonomy - in people, machines, and teams comprising both? In human beings, the notion is closely linked to that of responsibility - we are responsible for our actions only if we are autonomous (i.e. 'free') agents. Responsibility in turn is subject to two necessary conditions: if we are responsible for an action, then (i) we must know what action it is we are performing, and (ii) that action must be under our control (Coeckelbergh, 2020). Nevertheless, it may not be possible to supplement these two conditions in such a way as to yield jointly sufficient conditions (cf. Williamson, 2000) - perhaps autonomy has no definition, but is rather a family resemblance concept (Wittgenstein, 1953). This hypothesis is vindicated here by its fruitfulness in illuminating the applications that follow. In robotics, autonomous systems are characterised by an ability to act with independence and initiative. Independence means that the agent can perform a task without human supervision to achieve a goal, while initiative means that the agent is able to begin an action without a cue from a human. It is critical to know that these two qualities do not require or imbue intelligence in the machine, and are themselves orthogonal to each other and to intelligence, where intelligence is a quality that implies an ability to process information from the environment to answer a question. Embodied autonomous systems (e.g., robots) have tremendous potential to improve the human experience, if they can be developed from tools to true collaborative teammates. Teams are groups of 3 or more entities that work in concert though not necessarily in proximity toward a common goal. The interdependence of teammates improves the performance of the individual in homogeneous human teams, (e.g., Cooke & Lawless, 2021); however, performance of heterogeneous human-robot teams requires a level of mutual understanding that has yet to be achieved. We draw on results from philosophical simulations (cf. Mayo-Wilson & Zollman, 2021) of networks of inquiring agents to better understand the collective attitudes (List, 2014) that underpin many group phenomena (including teamwork). We then present a case study of human-machine interaction that exemplifies current pitfalls in machine understanding of human performance and the converse. In particular, we provide an overview of an experiment to enable high performing, "expert systems" to collaborate with humans as well as with non-self artificial intelligences, and discuss how the machines' learning of the game created a failure to understand (and follow) the rules driving human decisions in this high uncertainty task undermined the performance of the team. We conclude with an overview of what autonomous machines, as individual agents or working in teams (possibly comprising people), mean for human society and the types of interactions that must be supported to improve the human experience.

New College of the Humanities

Can XAI be 'Warfare Against Our Own Confusions'? Using Wittgenstein to Fulfill the Principle of Explainability

Zachary Goldberg (Trilateral Research)

29 July 16:10

The purpose of Explainable AI (XAI) is to use "the parameters relevant to decision-making to compute an account of the (algorithm's) output that is expressed in meaningful and explanatory terms" (O'Hara 2020, 1). AI regulations and guidance documents include explainability as an essential principle of ethical, responsible or trustworthy AI (e.g. AI Act in the EU; EU HLEG AI's Guidelines for Trustworthy AI; Alan Turing Institute's Understanding Artificial Intelligence Ethics and Safety). Meaningful explanations are especially important in socially relevant contexts employing AI such as medicine, defense, policing, education and finance. Years of research and technical progress have made significant strides towards developing explainable AI output. Nevertheless, fulfilling this principle remains a critical, ongoing challenge. How can reading Wittgenstein help us advance beyond the status quo towards creating more meaningful XAI? O.K. Bouwsma wrote that Wittgenstein's Philosophical Investigations (PI) "furnishes us with the rudiments of a certain warfare, instructs us in the use of certain instruments, instruments with which we are already furnished. Besides it furnishes us with exercises, exercises without end, war-games. And to what purpose? That we may perfect our skill in the warfare against our own confusions" (Bouwsma 1972, 82). This "skill" coincides with Wittgenstein's definition of understanding as "an ability to go on" (PI, §156). But how can XAI developers make sense of understanding as a "skill" or an "ability to go on"? And how does this interpretation help fulfill the goal(s) of XAI? Explainability is an instrumental value that furnishes AI-users with sufficient information for decision-making in a socially relevant content. One well-known XAI function is to explain how much an algorithm's input features contribute to its output in order for users to identify and mitigate proxies or bias leading to discriminatory actions. But simply knowing that particular inputs correlate with certain algorithmic output is not a full explanation (perhaps necessary, but not sufficient) within a socially relevant context. One cannot yet "go on" to use the algorithm in a non-discriminatory way. As Wittgenstein teaches us in PI, explanation/understanding is a social practice that involves an ability to go on, as well as the acceptance of this "going on" as an act of "going on". In the complex socially relevant contexts mentioned in the first paragraph, we can ask which actions one can go on to do and which actions would be socially accepted as reflecting an understanding of these complex contexts. The answer is that meaningful, socially aware understanding of complex contexts is the ability to go on to ask insightful ethical, social and political questions about issues that may not have settled answers—to wage warfare against our own confusions. As reflected during this conference and our philosophy classes, the socially accepted response to complex material is further inquiry rather than passive agreement. I will conclude this presentation with some practical suggestions for how XAI developers can facilitate the ability to go on to ask meaningful questions rather than producing static explanatory output. This Wittgensteinian approach will more readily meet AI regulations mandating explainability.

New College of the Humanities

Investigating Deepfakes: A Wittgensteinian Look on AI Regulation

Mehmet Taylan Cüyaz (Middle East Technical University)

29 July 16:50

Deepfakes are fabricated sorts of media created with deep learning techniques, which have produced a growing interest within philosophical literature. This paper aims to elucidate the discussion on deepfakes via investigating them under the Wittgensteinian notions. The paper is structured around three premises. The first premise intends to consider deepfakes as an instance of a language-game. Previously, culturally infused Wittgensteinian conceptual schemes were developed to understand the language used in technological contexts, and the term "technology games" was proposed. Yet, this premise addresses a different line of thought; instead of considering technology in a noun phrase, this paper proposes using it as an adjective to modify the concept of "game". Thus, technological games are a subset of ordinary language-games. Following this, different technological objects, as long as they serve as a means for communication, might have specialized language-games. Hence, being an instance of digital videos, deepfakes are strong candidates for having their language-game settings with unique rules. The second premise underlines a philosophical problem illuminated by a specific approach to deepfakes. "Epistemic Catastrophe Theses" focus on the destructive effects of deepfakes on social epistemology which depends on concepts such as knowledge, trust, and epistemic backstops. The problem lies not in overemphasizing on epistemic hazards. Rather, it is rooted in their analysis regarding the nature of video; which has a fixed character, the moving images depict reality with great accuracy. This stance echoes within the Tractarian fallacy; the term "video" is associated with an unchanging grammar that ensures a one-to-one correspondence within "what is the case". In this sense, the pursuit of an ideal or fixed grammar often leads to disorientation. The third premise bridges the gap exemplified in the second premise. A proper grammatical investigation of deepfake games needs to be considered in terms of proper restriction suggestions —videos exhibit both "depth" and "surface" grammatical properties. Depth grammar corresponds to the content of a video, and surface grammar refers to the spatiotemporal representations of pixels. Even if humans are masters of depth grammar, there are deepfakes with plausible content that are nevertheless fake. Therefore, grammatical investigation of the surface is a necessary condition to avoid epistemic challenges. Expert AI systems are far better masters of surface grammar since they are designed to detect minor clues at the pixel level. Therefore, the security of knowledge and trust in social contexts are cordially tied to the accuracy of our algorithms. Currently, the algorithms are deployed by social media platforms to safeguard the public from fake content. The detection mechanism regarding these forged media is operated through various tags-ticks, safe or unsafe marks – to inform/signal the end users. Herein, the companies act as insurers of knowledge or even as epistemic authorities. In conclusion, following a Wittgensteinian soul, this paper broadly argues that our questioning about the possible regulations must be changed from regulating the deepfakes to regulation of the emerging epistemic authorities.

New College of the Humanities

The Civic Status of Artificial Intelligence

Emanuele Bottazzi (LOA, IST-CNR)

29 July 17:30

According to David Strohmaier (2020) organisations are computing systems. We believe this model is not suitable for all types of organizations. Think of its oldest type, which has accompanied us humans for about 90% of our life on earth: the hunting party (Lee and Devore, 1968; Belardi J. B. et al. 2021). Computation here has at least a secondary role. On the other hand, it is true that there are organizations that have a functioning that could be assimilated to that of a computing system, as in states and other formal institutions. But this is also a mistake. From the scribe to the bureaucrat, whoever works within an institution to produce certain outputs from certain inputs finds herself within a form of life in which computing exists, as Wittgenstein goes, in mufti (in German: Im Zivil, RFM 257). It is the use outside computing that makes the computing an organizational activity in the institutional sense. As Peter Hacker (1993, 167) added, our calculating machines are always in uniform, never in mufti: computers save us the labor of calculating, but it is nonsensical to say that the machine infers or draws conclusions. We will try to show how this is strictly linked with Wittgenstein's reflections on following a rule in PI. As Lorenzo Bernasconi-Kohn (2007) poignantly resumed Martin Kusch's (2004) interpretation on this topic, we must avoid the categorical mistake «of assuming that conditionals that concern our normative practices are the same as conditionals about real or possible events». Moreover, we can illuminate the difference between actual organizations and computing systems by considering Wittgenstein's approach on contradictions. Much has been written about the beginning of PI's § 123: «A philosophical problem has the form: "I don't know my way about."». Much less has been written, instead, about his subsequent specification of this form at the end of PI's § 125, where the philosophical problem is « The civic (bürgerliche) status of a contradiction» and where this kind of contradiction « throws light on our concept of meaning something». At this level, the problem is not anymore technical, there is not and will never be a correct solution, but a solution that concerns what we want. This is not a way of shifting the problem, it is the very human activity of extricating not so much from the rules, but from the contrasting relations we have between our actions and what we want. We will try to show how these findings are of the utmost urgency in what, in some currents of sociology, is called algorithmic governmentality (Weiskopf and Hansen 2022), where organizations are gradually becoming computing systems in uniform. In this latter kind of organizations, the adoption of digital technologies points at reducing the noise of «unstructured, subjective opinion» (Bodie et al 2017, p. 964) that is, the human freedom of accepting, discussing and subverting the systems they live within.

New College of the Humanities

Turing's Philosophy of AI

Sebastian Sunday-Grève (Peking University)

30 July 14:00

The value of Turing's work on artificial intelligence has traditionally been reduced to what is now known as the Turing test, but it is more nuanced and compelling than previously assumed. Turing's thinking on this topic was far ahead of everyone else's, partly because he had discovered the fundamental principle of modern computing machinery, the stored-program design, as early as 1936 (a full twelve years before the first modern computer was actually engineered). Careful historical reconstruction of Turing's philosophy of AI shows that the heart of this work consists of logical investigations that proceed on the basis of what the later Wittgenstein called 'language games', which Turing employs in precisely the kind of function that Wittgenstein described as their being used as 'objects of comparison' (PI 130).

New College of the Humanities

Wittgenstein's influence on Turing's Test

Teresa Numerico (Roma Tre University)

30 July 14:40

There are many different interpretations of the Turing Test (see for example Shieber 2004, Moor 2006, Epstein, Robert, Beber 2009), but my aim here is to understand what Turing wanted to do by adopting the game as a possible discrimination between machines which would give themselves away and machines which could deceive a jury made of average non expert judges. I will argue that the formulation of the argument in favour of the substitution of the question about thinking machine and the imitation game was inspired by Wittgenstein's approach to language games and life forms. According to Turing the relevance of an intelligent task was related to the social and technical capabilities of the observer of attributing the intelligent behaviour to various devices or human beings (Turing 1948, Proudfoot 2020), position which is very similar to Wittgenstein's approach to language, to explain the process of understanding a sentence within a human conversation. This is the reason why the jury that is supposed to evaluate the success at the imitation game was mandatory made of non-expert judges and that it was necessary that the interviewing time was hypothesized as extremely short, around five minutes. The social capabilities that the machine needed to show was rather deceptive as well as the imitations of human behaviours were. If it is possible to interpret intelligence as a social characteristics attributed to an agent, no matter whether it was a human being or a machine, Turing was convinced that in some reasonably short period of time: "the use of the words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted" (Turing 1948, 449). It is likely that Turing was one of those happy few scholars who could perceive the novelty and the profound implications of Wittgenstein's research at that time of the lectures, even though he was exactly the negative target of Wittgenstein's arguments, and the discussions between them reproduced in the notes of the lectures, published posthumously (Wittgenstein 1939/1976), were rather tough during the meetings (see also Floyd 2016 on Turing-Wittgenstein relationships and reciprocal influences). The influence of Wittgenstein could be inferred also by looking at the content and at the form of the arguments adopted to reject some of the objections against the possibility of machine intelligence. In Turing 1950 well-known paper, the large section related to the confutation of the objections was very important and well organized. The objections were also used to promote the importance of the imitation game as a tool capable of facing many of the critical issues raised by detractors of machine intelligence. In some of the rejections of the objections Turing used Wittgenstein's argument structure: rejections of eventual counter-arguments to the defended thesis was precisely Wittgenstein's philosophical style. But the great philosopher seemed to have played a role also in suggesting the conversation as the main social device to assess intelligent behaviours of agents. Turing started the 1950 paper with the well-known question about "can machine think?". We will be surprised to see that this is the same question posed by Wittgenstein in the middle of the Blue Book, though in a different scenario: if it was clear that an amoeba certainly did not speak, or write or discuss, 'is it possible for a machine to think?' (Wittgenstein 1965, 47).

New College of the Humanities

The Turing Machine as a Boundary Object: Sorting out American Science and European Engineering

Edgar Daylight (Siegen University/Lille University)

30 July 15:20

We propose to consider the “Turing machine” as a boundary object in sociotechnical terms for two reasons.i First, computer scientists have defined the “Turing machine” in different ways. Some players characterized the machine with a one-way infinite tape, others preferred infinity in two directions. In both cases, the infinity involved has been taken to be an actual infinity by some and a potential infinity by others. Likewise, the workings of the machine have been defined with quadruple notation in certain books and with quintuple notation in others. Second, despite such mathematical variability, there is immutable content: each textbook definition adheres to the neo-Russellian tenet, that, everything a computer (or a physical object in general) can do, a Turing machine can do as well.ii The tenet conveys a computational version of logicism,iii which came to prominence in the 1950s with the writings of a first generation of computer scientists.iv The following 1958 words of John Carr illustrate this uptake of Turing machinery in connection with Artificial Intelligence: Based on Turing’s proof about universal machines:

1. Living organisms can be abstractly defined as symbol manipulator.
2. Actions ofliving beings can be described by a program.
3. Digital computers have all the features of Universal Turing Machines.
4. Digital computers can duplicate human beings.v

Our two observations, concerning the plasticity and neo-Russellian robustness of the Turing machine, adhere to the 1989 definition of a boundary object.vi Moreover, a boundary object has interpretive flexibility.vii And so does the Turing machine in the context of computer programming, as we shall illustrate in the next paragraph — and more elaborately in our presentation — with an American and a Dutch reception of the unsolvability of the halting problem (of Turing machines). Around 1967, Marvin Minsky and Edsger Dijkstra welcomed the neo-Russellian tenet that everything a computer can do, a Turing machine can do as well. Both men shared a common identity on opposite sides of the Atlantic Ocean, which benefited academic discipline building. Yet they positioned engineering in relation to Turing-machine theory differently. According to Minsky, mathematical theorems about Turing machines are rules that dictate the dos and don’ts of the engineer, including impossibility results (which hinge on an infinite abstraction of real machinery).viii While for Dijkstra, software engineering itself was a pure science. According to him, theoretical insights about Turing machines could, at best, advise the engineer.ix Examining the interpretive flexibility of the Turing machine as a boundary object sheds light on the contradistinction between American “science” and European “engineering,” and between natural laws and governing laws. The latter dichotomy was also contentious 30 years earlier, e.g., in an exchange between Alan Turing and Ludwig Wittgenstein in 1939, and remains so today in the philosophy of computer science.x Our historical findings will help software scholars compare intellectual cultures inside computer science, spanning multiple decades up till 2022.

i Susan Leigh Star, “This is Not a Boundary Object: Reflection on the Origin of a Concept,” Science, Technology, & Human Values 35, no. 5 (2010): 601.
ii Scott Aaronson, Quantum Computing Since Democritus (Cambridge: Cambridge University Press, 2013), xxviii; Michael Sipser, Introduction to the Theory of Computation (Boston: Thomson Course Technology, 2006), 137; David Harel, Algorithms: The Spirit of Computing (Reading, MA: Addison Wesley, 1992), 233.
iii Andrew David Irvine, "Bertrand Russell," The Stanford Encyclopedia of Philosophy (Summer 2017 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/sum2017/entries/russell/
iv Edgar Daylight, “Towards a Historical Notion of ‘Turing — the Father of Computer Science’,” History and Philosophy of Logic 36, no. 3 (2015): 205-228.
v John Carr, “Languages, logic, learning, and computers,” Computers and Automation 7 (1958): 21-22, 25-26.
vi Susan Leigh Star, James Griesemer, “Institutional ecology, ‘Translations’ and Boundary objects: Amateurs and Professionals in Berkeley’s Museum of Vertebrate Zoology,” Social Studies of Science 19, no. 3 (1989): 393.
vii Susan Leigh Star, “This is Not a Boundary Object: Reflection on the Origin of a Concept,” Science, Technology, & Human Values 35, no. 5 (2010): 601.
viii Marvin Minsky, Computation: Finite and Infinite Machines (Englewood Cliffs: Prentice Hall, 1967), 153.
ix Edsger Dijkstra, “EWD 316: A Short Introduction to the Art of Programming,” Technical report, November 1974, www.cs.utexas.edu/users/EWD/transcriptions/EWD03xx/EWD316.html; Edsger Dijkstra, “EWD 372: A simple axiomatic basis for programming language constructs,” Technical report, 8 May 1973, www.cs.utexas.edu/users/EWD/ewd03xx/EWD372.PDF; Edsger Dijkstra, “EWD 869: Ter afsluiting van de “Inleiding tot de Kunst van het Programmeren”,” Technical report, 7 Dec. 1983, www.cs.utexas.edu/users/EWD/transcriptions/EWD08xx/EWD869.html
x Edgar Daylight, “Addressing the Question “What is a Program Text?” via Turing Scholarship,” IEEE Annals of the History of Computing 43, no. 4 (October-December 2021): 87-91.

New College of the Humanities

Meaning as Use, Rule-bending, and AI safety

José Antonio Pérez Escobar (École Normale Supérieure Paris)

30 July 14:00

Imagine the following scenario. A “superhuman” AI is given a task—to reduce the number of people with cancer. One would expect that the AI would find new drugs, new treatments, and better means of diagnosis. We then notice that people start dying massively. It turns out that the AI poisoned running water in cities. Having fewer people means fewer people have cancer. Even worse—we try to stop it, but it tries to stop us from turning it off because it would be less capable of fulfilling its goal. What is more, even if we anticipated all this and tested the AI in a simulation first, it would realize that this was only in a simulation and would deceive us by “behaving properly”. I will explain recent concepts of AI safety, like mesa and base optimizers, the alignment problem (in its inner and outer version), and more. I claim that at the core of these problems there are rule-following issues that can be understood from a Wittgensteinian perspective. Late Wittgensteinian philosophy about language and mathematics, heavily focused on rule-following, may yield insights on how to develop AI safely. In principle, two strategies may be considered. The first is that we, humans, natural rule-benders, may stop bending rules by focusing on “clear” specifications of human goals. But even master programmers make mistakes and cause bugs. In this context, mistakes like these can be fatal. The second is designing rule-bending AIs. “Proper rule-following” is underdefined by rules, and clarification may help but this issue is in principle unsolvable. According to Wittgenstein’s notion of “meaning as use”, the best way to understand a rule is to see how it is used in natural contexts. AIs can be supervised and trained in human-like training contexts, in the spirit of the notion of rule-bending. I argue that this is a more feasible option. In this talk, I will unpack the second strategy, identifying the factors that, for Wittgenstein, lead to mutual understanding among humans. These factors are of a psychological, social, and cultural character. They must be included in AI training in order to “humanize” AI, in a way akin to how the later Wittgenstein conceived mathematics.

New College of the Humanities

Wittgenstein, Meaning, and the Chinese-Room: Intentionality as a Linguistic Phenomenon

David Chandler (King's College London)

30 July 14:40

Alongside topics such as rule-following and private language, some philosophers and scholars have often identified Ludwig Wittgenstein's consideration of the nature of meaning as one of, if not the perennial pursuits that occupied him in his later philosophy. Despite the obvious connection between the conception of meaning as use and intentionality, many are less than certain about how the two are said to interlace or interact with one another. In this paper, I want to address this interaction to ascertain the role of artificial intelligence within Wittgenstein's philosophy. In the writings that would become the 'Philosophical Investigations', Wittgenstein firmly denies the plausibility of thinking programmes, asserting that "machines can't think" (PI §360). And yet, some make use of Wittgenstein's writings to support the possibility of "strong" artificial intelligence—that is, those who believe that a program is, in fact, a mind. To remedy this confusion within our understanding, I shall approach the Chinese Room Experiment with this difficulty in mind. Let us suppose that a person is placed in a room to answer questions in Chinese. This person does not know any Chinese and therefore must rely exclusively on various manuals containing useful instructions for passing as a native speaker. Imagine that a machine answering these same questions passes the Turing test (Turing 1950). What this shows is that, as Searle explains, "instantiating a program could not be constitutive of intentionality, because it would be possible for an agent to instantiate the program and not have the right kind of intentionality" (Searle 1980, 450-51) However, as Obermeier observes, intentionality for Searle is a biological phenomenon. In contrast, for Wittgenstein, it is linguistic (Obermeier 1983, 340). As Wittgenstein states, "We use the words 'meaning', 'believing', 'intending' in such a way that they refer to certain acts, states of minds given certain circumstances" (BB 147). And again, "there are a great many combinations of actions and states of minds which we would call intending [...] For we could always have intended the opposite by reinterpreting the process of projection" (BB 32-33). This paper can be viewed as an attempt to explore the philosophical implications of intentionality as a linguistic rather than a biological phenomenon. I conclude by suggesting some limited remarks on Wittgenstein's metaphilosophical perspective both as a consequence of these discussions and as an explanatory factor.

New College of the Humanities

Wittgenstein, AI, Science and Technology

Roberto Presilla (LUMSA and Pontifical Gregorian University), Filippo Durisotto (Pontifical Gregorian University)

31 July 10:30

Wittgenstein's work on AI has been extensively studied from several points of view: some authors have underlined the pertinence of Wittgenstein's approach to AI, whereas others have pointed out criticalities residing in some parts of Philosophical Investigations and other works. The bare existence of research programs on AI suggests that AI is feasible. However an important question comes about: is AI a philosophical hypothesis, a scientific theory or a technological tool? The answer to this question involves many consequences, given that philosophy is definitely not science in Wittgenstein's view. The actual social context shows that the technological aspect is largely predominant. In the present work we analyse some timely applications of AI (especially in the field of machine-learning), aiming at determining their general impact. In particular we focus our attention on the difference between the algorithm's output and the deploy of the output itself, where the latter issue goes beyond the bare technological problem. Such difference is eminently theoretical, indeed is tightly bound to what we may call "reflexivity" or "self- consciousness". An analysis of argumentative strategies inside Philosophical Investigations shows that philosophical activity comprises "different methods" which are context-sensitive. We finally highlight some analogies between Wittgenstein and other philosophers.

New College of the Humanities

Wittgenstein's Woodsellers and Apparently Irrational AI

Mark Theunissen (Delft Technical University)

31 July 11:10

In the Remark on the Foundations of Mathematics (RFM 1.149) Wittgenstein provides the example of the "woodsellers", an apparently irrational people with the odd practice of measuring the value of stacks of wood by surface area rather than volume. This allows for the strange situation where an amount of wood that is large to us can be bought cheap if stacked high, whereas an amount of wood small to us can be much more expensive if laid flat. This thought-experiment about the intelligibility of apparently incommensurable arithmetic has long raised questions in the "rationality debates," the discussions in social science about how we should, or even whether we can, find seemingly irrational beliefs intelligible (e.g., Risjord 1993; 2001, Wilson 1970, Hollis and Lukes 1982, Theunissen, 2017). In this presentation, we argue how similar considerations are useful for thinking about how we should approach "black box" AI systems—typically called "agents"—who make decisions which surprise us and which seem irrational. After framing this problem in the context of current explainable AI debates, we canvas the different normative presuppositions of various interpretative and explanatory attempts to adjudicate apparently irrational practices: a principle of charity, an ad hoc explanatory approach, and a thicker pragmatic approach adjudicating competing interpretations by their measures of explanatory power. But applying these to examples of contemporary AI systems—such as chatbots, finance algorithms, and medical diagnostic systems—reveals their limits in helping us think of AI agents. This is because these approaches are non-neutral with respect to the sorts of agents whose beliefs or behaviors one tries to understand or explain—specifically, they assume the agents share a broadly human form of life. When applied to AI systems, this presumption that these machines are like us—that is, share in outline our forms of proper behavior and right action—threatens to distort what these AI are doing. We contend these accounts, at least in part, misses the radical nature of Wittgenstein's thought experiment. From an anthropological perspective, the thought-experiment not only brings into question whether these agents are rational; there is a further question whether interpreting these systems as rational agents, as subject to norms of reason altogether, is warranted. The typical approaches to the case of the woodsellers fail because these AI systems do not in any obvious way participate in our social practices, but their intelligibility to us is parasitic on those very practices. The upshot is that Wittgenstein's problem remains even more insoluble in the face of contemporary AI and requires novel approaches for figuring out how to treat AI agents which often act much as we do while remaining stubbornly uncanny for us.

New College of the Humanities

Can Machines Learn to Follow Rules? A Teleosemantic Approach

Jean-Charles Pelland (Bergen)

31 July 11:50

One of the most important components of AI is machine learning: we commonly talk about how algorithms can be trained to process large data sets and discover patterns among these, and we routinely hear of the ability of computers to learn from data to improve their performance in carrying out certain tasks. Without questioning how crucial machine learning is for the development of AGI, it is possible to ask whether what we call machine 'learning' can really count as a case of learning. While similar questions have been asked about the 'intelligence' of AI, the question of whether and how machines can genuinely be described as 'learners' has received comparatively little attention, perhaps because machine learning is a more recent construct. In this talk I explore to which extent we can genuinely describe machines as learners. The discussion will focus on questions related to what it means to learn to follow a rule, as presented in Kripke's (1982) discussion of Wittgenstein's rule-following paradox. Applying the distinction between implicit and explicit rule-following to data on how children learn the meaning of words for colours (Forbes & Plunkett 2018) and numbers (Carey 2009), I argue that talk of machines being able to learn is metaphorical, since artificial information-processing systems cannot be considered as following explicit rules - if they can be described as rule followers at all. In a nutshell, this is because machines, unlike children, cannot bridge the so-called "cognitive gap" (Loughlin 2014) between basic rule-following and the more explicit, content-driven rule-following that children come to master when learning languages. To illustrate the difference between these different forms of rule-following and how it could apply to information processing-systems, I survey criticism levied towards teleosemantic responses to Kripke's version of the rule-following paradox (e.g. Millikan 1990). Teleosemantic responses to Kripke have been criticized for relying on a metaphorical conception of rule- following that conflates co-variation with rule-following (Kusch 2006). Such criticism is based on the alleged under-determination of biological rules (Fodor 1990) and the alleged inability of biological norms to scale-up and provide the basis for explicit semantic norms (Burge 2010; Hutto & Myin 2014; Hutto & Satne 2015). After summarizing these worries about teleosemantics, I apply them to the case of machine learning. The idea here is that any criticism levied towards a teleosemantic explanation of rule following would have to apply to the information-processing system: if a hoverfly can't follow a rule, then why would we expect an information-processing system to do so? I end by showing that while teleosemantics might be able to use recent developmental data on core cognition (Carey 2009) to respond to its critics, there does not seem to be an equivalent response for machine learning, which suggests that machines cannot be considered genuine learners.

New College of the Humanities

Kripke on Wittgenstein About Intention and Correctness: A Sceptical Paradox for Computation

Chiara Manganini (University of Milan)

31 July 10:30

According to the exegesis offered by Kripke (1981), to the heart of the Philosophical Investigations (Wittgenstein 1953) we find the refutation of a mentalistic conception of rule-following. Kripke believes that Wittgenstein's dismissal takes the shape of a sceptical paradox for the metaphysics of language, as the inexistence of intentions ultimately forces us to conclude that no word of language is governed by a usage rule. In this paper, I argue that the pervasiveness of the scepticism about rule-following ends up affecting another portion of metaphysics as well, namely that of computation. The idea of a computing system following a rule is captured by the central philosophical notion of physical implementation, which still today is at the centre of contemporary debate. For my part, I will focus on a specific theory of implementation called "ontology of levels of abstraction" (or "LoAs ontology") which, quite unusually, endorses the very Wittgensteinian observation that no physical fact-of-the- matter about a machine can univocally determine which function it is implementing. The proponent of such a view, in fact, believes that the question of which function is being physically implemented by a computational system can only be determined by the content of a certain specific human's intention – namely, that of the human who programmed the system itself. In this respect, the LoAs ontology and Wittgenstein's late philosophy cannot differ more: indeed, within the framework of LoAs ontology, Wittgenstein's metaphysical scepticism about intention would lead to the unacceptable conclusion that nothing can determine which abstract computation a machine is implementing and, therefore, whether its output is correct or incorrect. I will take this to be a sceptical paradox analogous to the one Kripke attributes to Wittgenstein, but this time concerning the metaphysics of computation. I will then take into consideration another sceptical claim that has been troubling the philosophy of computation for quite some time, i.e., the Triviality Argument; I will explore the possibility of solving the sceptical paradox for computation by exploiting the strategy originally devised by Chalmers against Putnam's Triviality Argument. I will hence attempt to sketch a sceptical solution for the paradox, along the lines of the one Kripke attributes to Wittgenstein. I will finally highlight that this approach to the question of physical implementation owes a lot to some pragmatist theses that, starting from the late 70s, have been developed on the limits of the methodology of program verification.

New College of the Humanities

Law, Rules, Machines

Gianmarco Gori (Vrije Universiteit Brussel)

31 July 11:10

The concept of rule plays a key role in law as much as in Artificial Intelligence. I argue that a certain picture of rules holds jurists, computer scientists and data scientists captives and that the conceptual confusion generated by such picture plays a foundational role in enabling the bridging of law and AI. This conceptual confusion is productive in that it affords the design and use of computational legal machines. As mathematicians with calculating machines, jurists can exploit the convergence between the causal mechanisms governing machines hardware, the rules-calculi of programming languages and legal rules to ascribe normative significance to machine operations. At the same time, the success of such conceptual confusion risks further abstracting rules from normative practices. While the code-driven approach of GOFAI reiterates the conceptual muddling addressed by Wittgenstein in his discussion of rules as calculi, the data-driven paradigm characteristic of the current AI spring hinges on the blurring of the lines between rule-like behaviour and rule following. The picture of rules assumed by AI research is particularly congenial to jurists. As the metaphor of law as a machine represents a recurrent topos in the history of legal thinking, jurists have been seeking rules as hard as the material with which the mechanisms of such machine are made. However, narrow conceptions of legal rules have persistently proved a short blanket: not only they fail to account for how rules change "as we go along", they also undermine their own foundations by running into a problem of infinite regress. AI offers to jurists a conceptual toolbox to account for such aporias without undermining the narrow picture of legal rules. Rules fail because human rule followers are defective carbon-based machines. Their flaws are apparent when compared to "silicon-based machines", which are instead presented as "rule following beasts". At the same time, the success achieved by ML research in the field of pattern detection and prediction gives new impetus to behaviourist legal perspectives: as the practice of jurists is a "just very much more complicated" manifestation of rule-like behaviour, AI can detect the "real rules" which drive legal actors, dismissing the reasons and justifications that jurists articulate. I claim that the conceptual confusion can be dissolved through a closer look at the role that the concept of Verständigung plays in Wittgenstein's remarks on rule following. I argue that the attention paid to the issue of explainability within the current debate on AI and law is misplaced to the extent that it invites reflection on rule followers in terms of "how they are done" instead of "what they do" and, more importantly, "what it is that which they do". Addressing the latter questions enables to relocate the discussion in the perspective of the criteria which license the ascription of meaning and govern the assessment of what counts as following a rule within the practice of jurists.

New College of the Humanities

Black Boxes, Beetles and Beasts

Ian Ground (University of Newcastle)

31 July 14:00

A common ethical objection to certain classes of A.I system – the Black Box issue – depends on the realisation that answers to the question of why (at least for some senses of "why"), the system made a particular "decision" are, logically, unavailable. This objection is, or should be, independent of the objection that an A.I system may be trained on data which is some way unjustly skewed. The implications of the Black Box issue for the philosophy of mind and in particular our conception of normative rationality are less frequently explored. Can a decision be regarded as rationally based if, in principle, it is not possible to track that decision back through a series of deductions, inferences, and principles? Is it legitimate to compare such cases to our everyday reliance on testimony? What are the implications for the conceptual possibility of minded machines? In these debates, Wittgensteinians may find themselves facing some dilemmas. Many are, at least temperamentally, inclined to scepticism regarding claims about the putative intelligence of machines, rejecting a range of assumptions upon which those claims depend. However, for classes of A.I to which the black box problem applies, many of those assumptions are not in play. Moreover, Wittgenstein's remarks on rule-following reject the idea that rational cognition is possible only via encoded representations of rules. The same rejection is central to the positive case for machine intelligence in the A.I case. How then should the Wittgensteinian respond to these issues? In this discussion, I offer reasons for thinking that Wittgensteinians should be intensely relaxed about such A.I systems. It remains for the Wittgensteinian to resist the conceptual possibility of machine intelligence by insisting on the foundational role of the concepts of biological life and (en-) action perhaps in tandem with a quietist attitude towards explanation. The discussion concludes by raising some concerns about this strategy via a comparison with the case of non-linguistic animals.

New College of the Humanities

The Beetle in the (Black) box. A Wittgensteinian Guide to Machine Learning

Giovanni Galli (University of Urbino)

31 July 14:40

In the last years, we witnessed remarkable progress in the development of Natural Language Processing (NLP) systems, which are neural networks implemented with self-supervised deep learning methodologies which learn concept representation from rough data and are very effective in targeting tasks such as question answering, textual entailment, and translation (Devlin et al., 2019; Kitaev et al., 2019, Wang et al., 2020). As in other relevant research topics, insofar as deep learning technologies give impressive effective results, the system's output comes out from what is usually labelled the “black box”. This lack of clarity concerning the processes involved in the deep learning flow lead to what is called now Explainable Artificial Intelligence, a movement working on the AI that can explain its decisions to humans. Moreover, it is well known how Wittgenstein’s works contributed to the progress of NLP. Famous and crucial concepts such as Satzsysteme and language-game, Sprachspiele, or meaning-as-use are still implied at the basis of NLP systems. There is, indeed, a subsidiary concept that I want to analyse here, which is the private-language argument, expressed via the beetle-in-the-box argument (§293 Philosophical Investigation), with specific importance for deep learning methodologies used in language processing systems. Skelac and Jandrić (2020) analysed the concept of meaning-as-use from Wittgenstein and Firth to Google’s Word2vec, a word vector representation developed by Google in a machine translation model. They made an important work that emphasizes the relevance of context concerning word meaning in Wittgenstein, Firth, and Word2vec, and underlies the scope differences of context in each case. As we know, the Beetle in the box argument shows the rejection of a private language construed only through an internal process the individual of the thought experiment calls the “beetle”, but which does not refer to humans as an actual “beetle” would. First I want to answer whether Wittgenstein’s beetle-in-the-box argument is relevant also to the case of the black-box property in deep learning systems. I want to figure out what are the implications for the further development of NLP. Second, it is important to analyse the role of the concept language game in the NLP systems, how it is embedded and why. According to this question, the first to notice is that there is a correlation between Wittgenstein's invitation to look, which is an invitation to dismiss the aim of theorizing about languages, and the absence of theory- ladenness in deep learning technologies involved in NLP. This kind of research context is called by Pietsch (2022) phenomenological science, distinct from theoretical science. According to him, phenomenological sciences make use of deep learning technologies without a hard theoretical burden, which is a feature of a specific field of data science characterized by an inductive methodology. It holds also in the case of NLP systems. Third, I argue that even if we can distinguish a strong and a weak definition of a private language, Wittgenstein’s argument holds also for these models and his worries are still a good guide for NLP developers.

New College of the Humanities

Is 'Accuracy' Accuracy?

Chad Lee-Stronach (Northeastern University)

31 July 15:20

Artificial intelligence is being used to automate decision-making in many of the social institutions that shape our lives. It is often based on models that use machine learning algorithms to optimise an objective function, based on data that putatively represent the normatively-relevant features of the context of decision-making. This optimisation process is typically invoked to ground claims about the accuracy of these systems. Accuracy, in turn, is invoked to justify the deployment of these systems. In this talk, I argue that many, perhaps all, existing applications of artificial intelligence are methodologically unjustified. Automated systems are, explicitly or otherwise, based on normative models that purport to accurately represent the normatively relevant considerations of the contexts to which they are de- ployed. But what is an 'accurate' normative representation and what does it mean for these artificial intelligence systems to be accurate in this sense? I explore the conceptual problem of normative representation by reference to early Wittgenstein's picture theory of representation.1 A central lesson from this analysis is that models – no matter how artificially intelligent – cannot judge or ensure their own representational accuracy in normative domains. From this, I identify general lessons for emerging discussions of the methodology of normative modelling.2 I then apply these lessons to recent technical approaches that purport to 'interpret', 'explain' or otherwise justify the decisions of artificial intelligence in social institutional contexts.3

1. Ludwig Wittgenstein, Tractatus Logico-Philosophicus (London: Routledge Kegan Paul, 1922).
2. Franz Dietrich and Christian List, "What Matters and How it Matters: A Choice- Theoretic Representation of Moral Theories," The Philosophical Review 126, no. 4 (2017): 421–479; Michael Weisberg, Simulation and Similarity: Using Models to Understand the World (Oxford: Oxford University Press, 2013); Michael G. Titelbaum, "Normative Mod- eling"
3. Cynthia Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence 1, no. 5 (2019): 206–215, arXiv: 1811.10154.

New College of the Humanities

Can Machines Think? A Wittgensteinian Approach to the Fundamental Problem of the Philosophy of AI

Arturo Vázquez (University of Southampton)

31 July 14:00

This paper shows how a Wittgensteinian approach can contribute to clarifying and solving the fundamental problem of the philosophy of artificial intelligence (AI), namely, whether or not it is possible to build thinking machines. In contemporary philosophical discussions of AI, authors generally address such a problem from ontological, epistemological, or technical standpoints (see e.g. Chalmers 2016, Pinker 1997, and Dreyfus 1972). Only a Wittgensteinian approach can take the required previous step and clarify the philosophical assumptions in the language in which this problem is formulated first (PI, 109). The first part of the paper offers a general reconstruction of the fundamental problem in the philosophy of AI, based on the distinction between the strong and weak conceptions of artificial intelligence (Searle 1980). After, we classify some of the authors that engage in these debates correspondingly. Part two explores some of the Wittgensteinian approaches to this problem to date, and argues that these efforts are located within, and represent manifestations of, the weak conception of AI (see Harre 1988, Casey 1988, Neumaier 1989, Emiliani 1990, and Hacker 2019). In general, the authors that apply Wittgensteinian methodologies to address the fundamental problem in the philosophy of AI understand the grammar of everyday language as the standard against which we can determine whether certain innovative language uses and expressions make sense. The skeptical attitude displayed by most Wittgensteinian approaches to the problem of the applicability of psychological predicates to machines is prima facie natural. However, the third part of the paper argues that one problem in these approaches is that their authors seem not to appreciate enough the fact that the grammar of our language is dynamic (OC 96-7. See Klagge 2017). A critical assessment of the former philosophical perspectives and the analysis of the application of psychological predicates to machines in AI (Silver et al. 2018) help further develop a Wittgensteinian account of the philosophical problem of AI by arguing that such a problem expresses a false dichotomy that requires a non-binary answer. In turn, this paper represents an attempt to articulate different Wittgensteinian perspectives with a larger canvas of more mainstream, contemporary philosophical concerns about these issues.

New College of the Humanities

Could a Machine Think? Wittgensteinian Perspectives on Computing Machinery and Intelligence.

Samuel Pedziwiatr (LMU Munich)

31 July 14:40

Could a machine think? Wittgenstein repeatedly returns to this question in his later writings to investigate the grammar of "thinking," for example in the Big Typescript, the Blue Book, and in §359 of Philosophical Investigations. The development of Wittgenstein's remarks is reflective of the subtle changes in his philosophical method and of his evolving views on the relation between mental phenomena, phenomenal experiences, and the subject who has them. In his article "Computing Machinery and Intelligence," Alan Turing argues that the question whether machines can think is meaningless, and proposes to replace it by considering a language-game scenario instead: the imitation game. To which extent does Turing engage with Wittgenstein in his article? In my talk, I wish to address this question by reassessing the available textual and archival evidence and by critically comparing Turing's and Wittgenstein's arguments. Special attention is paid to the historical background and argumentative context of Wittgenstein's remarks in the Nachlass. When Turing published his paper in Mind in 1950, Norman Malcolm wrote Wittgenstein to ask whether Turing had attended his lectures on the foundations of mathematics and whether the article was a hoax. Wittgenstein responded on 1 December 1950: "You're quite right, a mathematician by the name of Turing attended my lectures in '39 (they were pretty poor!) & it's probably the same man who wrote the article you mention. I haven't read it but I imagine it's no leg-pull." Juliet Floyd has argued that the anti-psychologistic outlook and emphasis on the social aspects of ordinary language in "Computing Machinery and Intelligence" can be seen as distinctively Wittgensteinian features. In her view, Turing's philosophical tenets were profoundly shaped by personal discussions with Wittgenstein and Alister Watson, as well as by Turing's familiarity with Wittgenstein's lectures and writings. This narrative has been challenged by other Wittgenstein scholars such as Ray Monk, who have maintained that Turing's views on the philosophy of mind, logic, and mathematics are antithetical to Wittgenstein's anthropological perspective. My talk explores a middle ground view. The alternative reading of Turing's engagement with Wittgenstein that I wish to propose is that the opening sections of "Computing Machinery and Intelligence" can be interpreted as a grammatical analysis in a (somewhat) Wittgensteinian vein. Turing attacks the question "can machines think?" in a way that resembles Wittgenstein's treatment of the question "what is the meaning of a word?" in the Blue Book: he considers an analogous problem that avoids the potential conceptual confusions and metaphysical baggage of the question in its original formulation. At the same time, Turing is highly critical of the Wittgensteinian notion that philosophical problems can be solved by examining the common, everyday usage of words. On this reading, "Computational Machinery and Intelligence" can be seen as Turing's attempt to meet philosophers' potential objections to the idea that it is possible to construct thinking machines on their own ground.

New College of the Humanities

Between Wittgenstein and Turing: Enactive Embodied Thinking Machines

Tomi Kokkonen (University of Helsinki), Ilmari Hirvonen (University of Helsinki)

31 July 15:20

In the Remark on the Foundations of Mathematics (RFM 1.149) Wittgenstein provides the example of the "woodsellers", an apparently irrational people with the odd practice of measuring the value of stacks of wood by surface area rather than volume. This allows for the strange situation where an amount of wood that is large to us can be bought cheap if stacked high, whereas an amount of wood small to us can be much more expensive if laid flat. This thought-experiment about the intelligibility of apparently incommensurable arithmetic has long raised questions in the "rationality debates," the discussions in social science about how we should, or even whether we can, find seemingly irrational beliefs intelligible (e.g., Risjord 1993; 2001, Wilson 1970, Hollis and Lukes 1982, Theunissen, 2017). In this presentation, we argue how similar considerations are useful for thinking about how we should approach "black box" AI systems—typically called "agents"—who make decisions which surprise us and which seem irrational. After framing this problem in the context of current explainable AI debates, we canvas the different normative presuppositions of various interpretative and explanatory attempts to adjudicate apparently irrational practices: a principle of charity, an ad hoc explanatory approach, and a thicker pragmatic approach adjudicating competing interpretations by their measures of explanatory power. But applying these to examples of contemporary AI systems—such as chatbots, finance algorithms, and medical diagnostic systems—reveals their limits in helping us think of AI agents. This is because these approaches are non-neutral with respect to the sorts of agents whose beliefs or behaviors one tries to understand or explain—specifically, they assume the agents share a broadly human form of life. When applied to AI systems, this presumption that these machines are like us—that is, share in outline our forms of proper behavior and right action—threatens to distort what these AI are doing. We contend these accounts, at least in part, misses the radical nature of Wittgenstein's thought experiment. From an anthropological perspective, the thought-experiment not only brings into question whether these agents are rational; there is a further question whether interpreting these systems as rational agents, as subject to norms of reason altogether, is warranted. The typical approaches to the case of the woodsellers fail because these AI systems do not in any obvious way participate in our social practices, but their intelligibility to us is parasitic on those very practices. The upshot is that Wittgenstein's problem remains even more insoluble in the face of contemporary AI and requires novel approaches for figuring out how to treat AI agents which often act much as we do while remaining stubbornly uncanny for us.