L'HUMAIN EST-IL UNE MACHINE, OU EST-IL AUSSI ESPRIT ?*

FRANÇOIS FLEURET

Juin 2015

Le développement des ordinateurs et l'augmentation constante de leur complexité et de leur puissance de calcul depuis plus de six décennies ont réduit l'écart entre leurs "capacités cognitives" et celles des humains.

Cette évolution rend la différence entre les humains et les machines de plus en plus floue. En particulier elle rend caducs nombre d'arguments qui reposaient sur l'évidente trivialité du "comportement" des machines.

1 INTRODUCTION

Quelle est la réalité du substrat de l'esprit humain ? Est-il une très complexe horloge, ou bien implique-t-il "quelque chose de plus" ? C'est une question nuancée et difficile, à laquelle pourtant la grande majorité des gens répondent avec certitude par la négative : Non, l'humain n'est pas qu'une "machine". Comme le titre de cet essai le formule, il est aussi "esprit".

Comme il a été difficile d'accepter que l'homme est le descendant d'animaux, ou que la vie émerge de la matière inerte, il est choquant aujourd'hui d'imaginer que le psyché, les sentiments et la conscience de soi pourraient émerger d'opérations électriques ou mécaniques. Cette opinion provient probablement d'une pré-conception fausse de ce que sont les machines, et implicitement de leurs limitations.

*Cet essai reprend les arguments d'un débat avec le Prof. Wulfram Gerstner organisé par l'association Cèdres Réflexion le 15 juin 2015 à Lausanne.

2 LA NATURE DE LA QUESTION

La difficulté de la question vient en partie de la complexité des trois termes qu'elle combine. Si "humain", "machine" et "esprit" ont des significations à peu prés partagées dans les discussions de tous les jours, elles ne sont pas suffisantes pour analyser la question que nous nous posons.

2.1 *Machine*

Au sens courant une *machine* est un dispositif construit par l'homme qui peut utiliser une source d'énergie pour effectuer une tâche sans intervention humaine : un moulin, une machine à laver, ou une horloge.

Le sens de "dispositif" n'est pas clair, et la contrainte qu'il doit être "construit par l'homme" est excessive. On accepte parfaitement que des machines sont fabriquées par d'autres machines. Si c'est de la conception et non de la construction dont on parle, ce n'est pas clair non plus. De nombreuses machines – avions, moteurs, microprocesseurs – sont aujourd'hui conçues en partie par des ordinateurs qui calculent des formes et des configurations optimales.

Pour la discussion qui nous occupe, nous devons en revanche imposer des contraintes sur les composants que nous acceptons dans une machine. Il est probable que l'on n'appellerait pas "machine" l'ensemble composé d'un moulin et de l'animal qui le fait tourner. Nous ne pouvons pas accepter que des structures complexes soient prises "dans la nature" et utilisées telles quelles. Un tel hybride pourrait en particulier inclure des fragments de cerveaux humains et la question se viderait de sens.

*
* *

Il semble que pour la plupart des gens les machines sont limitées parce qu'elles sont "totalement réductibles", c'est à dire qu'elles sont des assemblages d'un nombre fini de composants élémentaires, chacun ayant un comportement parfaitement spécifié, déterministe, et simple.

Pour la suite, nous prendrons comme définition d'une machine que c'est un dispositif qui a un état interne qui peut prendre un nombre fini de configurations, et qui évolue par pas de temps de manière parfaitement spécifiée. Sans influence extérieure, son état à l'instant t détermine complètement son état à l'instant $t+1$. Une telle définition correspond à un objet mathématique précis : la machine de Turing (Turing, 1936), et modélise en pratique ce qu'est un ordinateur.

2.2 *Humain*

Nous ne nous intéressons évidemment pas à l'apparence ou à la constitution physique de l'être humain, mais à ses capacités cognitives "au sens large". C'est à dire à son aptitude à résoudre des problèmes dans le monde réel – qu'ils soient intellectuels ou moteurs – mais aussi à avoir une expérience consciente.

La difficulté vient de ce dernier point. Nous avons chacun un *impression subjective d'être*. Mais nous ne pouvons pas qualifier l'humanité d'autrui selon *son* expérience subjective. Nous ne pouvons, ni en pratique, *ni en principe,* être à la place d'autrui. Nous pouvons imaginer ce qu'il ressent en utilisant notre propre expérience comme modèle, mais nous n'avons aucun moyen de savoir si c'est *son* expérience.

L'argument est essentiellement que même si vous aviez accès au fonctionnement interne du cerveau d'une autre personne, vous devriez disposer exactement des mêmes moyens qu'elle pour interpréter ledit fonctionnement interne, donc cette interprétation devrait être faite par *son* cerveau. Tant qu'une partie du traitement est faite par le vôtre, l'interprétation qui en résulte n'a aucune raison d'être correcte. Mais si aucune fraction de votre cerveau n'est impliquée, *vous* ne ressentez pas les choses, *l'autre* les ressent (Nagel, 1974; Dennett, 1991).

*
* *

La seule alternative pour définir ce qu'est un humain est donc une caractérisation extérieure. Ne pas le définir selon son expérience subjective, à laquelle rien ni personne d'autre n'aura jamais accès, mais selon son comportement dans le monde.

Une telle caractérisation prend la forme d'un test comportemental tel que le test de Turing (Turing, 1950). Le principe de ce dernier repose sur une épreuve au cours de laquelle un examinateur humain communique avec un interlocuteur qu'il ne voit ni n'entend. L'interaction se faisant donc par exemple à l'aide d'un clavier et d'un écran. Cet examinateur doit deviner à l'issue d'une discussion avec son interlocuteur si ce dernier est un humain ou une machine qui "fait semblant" d'être un humain. Si la machine arrive à tromper des examinateurs plus de la moitié du temps au cours de telles épreuves, on estime qu'elle passe le test.

Décider si ce test est assez contraint – combien de fois la machine doit elle passer l'épreuve ? quelle est l'intelligence et l'expertise des examinateurs ? – et s'il est suffisant est un problème philosophique que l'on ne résoudra pas ici (Dennett, 1995; Chalmers, 1996).

Néanmoins, il ne faut pas sous-estimer la richesse d'un tel test. La discussion avec la machine peut aborder des questions liées aux émotions, à la conscience de soi, et

aux processus introspectifs qui lui sont associés. Pour le passer, la machine doit donc au moins savoir *imiter* les capacités introspectives et l'aptitude à rapporter une expérience subjective cohérente. De fait, nous considérons comme acquise l'humanité des personnes avec qui nous interagissons tous les jours, alors que nous ne les connaissons qu'à travers une version limitée du test de Turing.

2.3 *Esprit*

Le terme "esprit" est probablement le plus mal défini des trois qui nous intéressent. Il peut aussi bien faire référence aux fonctions cognitives qu'à des composantes incorporelles ou divines de l'humain. Le sens implicite dans la discussion présente est la partie non-physique de l'humain, donc "ce qui n'est pas une machine".

Cette interprétation est dangereuse car elle implique qu'une machine n'existe que dans la réalité physique. Or, une importante partie de ce qui fait l'utilité des machines modernes, en particulier les programmes informatiques, sont de l'information pure, indépendante du substrat dans lequel elle est représentée. Elle peut être dupliquée, transmise et encodée sous des formes diverses. Le même calcul peut être réalisé à l'aide d'un dispositif électronique, mécanique ou hydraulique, ou par un humain avec un crayon et des feuilles de papier.

Une définition de l'esprit – qui ne serait pas en opposition à "machine" – pourrait donc être que c'est le modèle "informationnel" d'un système cognitif, détaché de sa réalisation physique, comme une partition de musique ou une recette de cuisine sont des représentations qui peuvent se matérialiser en objets tangibles. Il est donc immatériel et permanent.

2.4 *Quelle est la question*

Ces définitions étant posées, la question devient donc : "Peut-il exister un ordinateur qui passerait le test de Turing ?" ou "Un humain peut-il faire quelque chose qu'un ordinateur ne peut pas faire *en principe* ?"

La conviction que l'humain peut faire plus qu'une machine vient probablement d'une idée fausse de ce qu'est une machine, elle-même due à l'observation des machines de la vie courante. Ces dernières ne représentent qu'une infime partie de l'ensemble des machines qui peuvent exister en principe, et induisent une intuition erronée à leur sujet.

3 LA RÉALITÉ DES MACHINES

Les machines sont perçues comme étant de complexité limitée et incapables de sortir des comportements spécifiés par leurs concepteurs. Les ordinateurs modernes forcent à reconsidérer cette vision.

3.1 *Complexité*

Le terme "machine" recouvre des objets aussi divers qu'un moulin à vent, un métier à tisser, ou un ordinateur. Or les complexités de ces différents dispositifs sont d'ordres totalement différents.

Un ordinateur standard, comme on peut l'acheter dans une grande surface, est équipé d'une mémoire qui contient plusieurs dizaines de milliards de chiffres, et d'une unité de calcul capable de faire plusieurs dizaines de milliards d'opérations par seconde. Ces valeurs ne sont pas des grandeurs hyperboliques à la signification obscure, mais des spécifications concrètes : Un ordinateur qui coûte quelques centaines de francs fait des multiplications cent fois plus vite que l'humanité *dans son ensemble.*

Cette immense complexité rend confuse la nature du processus qui prend corps dans les calculs effectués par un ordinateur.

*
* *

Un grand nombre de techniques de simulations, par exemple pour les prédictions météorologiques, l'optimisation de la résistance des habitacles de voitures, ou la conception de fuselages d'avions, reposent sur la même idée "d'éléments finis", qui consiste à décomposer l'objet d'intérêt en toutes petites parties. La motivation derrière cette approche est que l'on sait comment se comporte un petit volume d'atmosphère ou de métal lorsqu'il subit des contraintes, mais on ne sait pas quel est le comportement global qui en découle. Un ordinateur, grâce à sa puissance de calcul, peut mener le même calcul "simple" des milliards de fois, "petite partie par petite partie", et finalement simuler un comportement global. Cela permet de répondre à des questions à propos de ce comportement global : "est-ce que l'habitacle s'écrase trop et met les passagers en danger ?" ou "est-ce qu'avec cette forme d'aile, l'avion consomme plus de fuel ?"

Lorsqu'une telle simulation fonctionne dans un ordinateur, il en résulte deux niveaux de réalité différents : celui où se fait le calcul proprement dit, qui est une succession d'un très grand nombre de calculs simples, et celui où existe

l'objet d'intérêt. Si nous simulons une horloge mécanique dans un ordinateur en traitant des petits bouts de matière, et que "les aiguilles tournent", ce qui veut dire concrètement que les chiffres dans la mémoire de l'ordinateur fluctuent d'une manière qui peut être interprétée comme cela, de quoi parle-t-on exactement ? Est-ce que ces aiguilles "existent" ?

Si l'on simulait un cerveau humain dans un ordinateur, molécule par molécule, en quoi notre compréhension des mécanismes de simulation nous donnerait une intuition valide à propos du résultat de cette simulation ?

3.2 *Prédictibilité et optimisation*

La seconde limitation supposée des machines est leur prédictibilité.

Un argument trivial contre cette idée est que l'on peut faire des machines non-déterministes. Au sens courant en utilisant des processus mécaniques chaotiques comme on le fait pour une machine de loterie, ou bien dans un sens plus exact en utilisant des mesures de phénomènes quantiques.

Cette astuce n'est pas convaincante, mais elle permet de comprendre que ce n'est pas un déterminisme au sens strict que l'on reproche à une machine, mais plutôt un "périmètre d'existence", d'où nous pensons qu'elle ne pourra jamais s'échapper. Nous savons quels sont le passé et le futur d'une horloge, nous savons quelle trajectoire elle va suivre pour toujours.

Un tel périmètre est beaucoup plus difficile à définir pour un programme informatique car ce dernier ne suit pas une série d'opérations qui sont exécutées les unes après les autres de manière systématique. Il décide quelles opérations effectuer en fonction de celles qu'il a exécutées précédemment. Ce comportement "conditionnel" est complètement spécifié par le programmeur, mais il en résulte une complexité de fonctionnement bien supérieure à la complexité du programme lui-même.

*
* *

Combiné à l'immense puissance de calcul des ordinateurs, ce comportement permet à un programme informatique de produire en pratique un résultat qui va au-delà de ce que ses concepteurs peuvent anticiper. Un programme d'échec peut trouver un coup qu'aucun humain n'aurait pu imaginer, et un système de contrôle de bras robot peut trouver une manière de positionner une pièce dans un assemblage en usine plus efficace que la meilleure stratégie imaginée par les humains avant lui.

Plus généralement, la plupart des programmes d'optimisation – dont un programme d'échec ou le contrôleur d'un bras robot sont deux exemples – n'ont d'intérêt précisément que parce que ce qu'ils produisent n'aurait pas pu être trouvé par un humain. Ils fournissent des solutions "meilleures" simplement parce qu'ils ont à leur disposition une puissance de calcul et une capacité mémoire supérieure à leurs concepteurs. Par principe, leur utilité est directement liée à leur imprévisibilité.

3.3 *Apprentissage statistique*

Dans tous les exemples qui précèdent, une distinction claire persiste entre les programmes, conçus par les humains, et les résultats que ces programmes produisent. Or, cette distinction est arbitraire.

Il arrive très fréquemment que le fonctionnement d'un programme dépende de paramètres qui sont adaptés empiriquement en fonction des données qu'il doit traiter. Cette adaptation est très proche du réglage mécanique d'une machine pour l'adapter au contexte de son utilisation.

Un important domaine de l'informatique contemporaine, l'apprentissage statistique, étudie et développe des méthodes qui permettent d'adapter de très grands nombres de paramètres. Ce type d'approche est particulièrement utile pour faire de la *prédiction* pour laquelle il est difficile de définir une règle formellement. Par exemple pour prédire automatiquement quel objet est visible sur une image, pour déterminer si une cellule est cancéreuse à partir d'une signature d'expressions de gènes, ou pour reconnaître un mot dans un enregistrement sonore. Ces tâches pourtant très faciles pour les humains et les animaux quand il s'agit de vision et d'audition ont été jusqu'à ces dernières années le talon d'Achille des ordinateurs.

Le principe de la méthode la plus utilisée pour ces tâches consiste à modifier progressivement les paramètres pour que le programme prédise ce qu'il doit prédire sur des "exemples d'apprentissage", par exemple des images pour lesquelles on a indiqué quels objets sont visibles. Pour chacun de ces exemples, l'ordinateur apporte une petite correction à chaque paramètre pour que la réponse prédite soit "plus proche" de la réponse désirée (Rumelhart et al., 1986).

*
* *

La complexité des ordinateurs, combinée à ce type de méthodes d'apprentissages, rend la distinction entre un programme et le produit d'un programme très floue.

Le nombre de paramètres que de telles techniques peuvent estimer atteint aujourd'hui plusieurs milliards, et leur permet d'avoir des propriétés d'universalité. En pratique, tant qu'il lui est fourni assez d'exemples d'apprentissage, un tel programme apprendra comment faire une prédiction correcte, quelle que soit la complexité de la règle correspondante.

Pour certaines applications réelles comme l'analyse du langage ou la sélection de molécules médicamenteuses, des équipes sans expertise du domaine ont obtenu des résultats meilleurs en utilisant ces techniques d'apprentissage que des experts du domaine en construisant des modèles explicitement (Collobert et al., 2011; Ma et al., 2015). On arrive donc dans un tel cas à une situation où une forme de "compréhension du monde" provient d'un calcul et de données d'apprentissage, et en aucun cas des concepteurs humains du programme.

De plus, si les structures obtenues par apprentissage font bien ce qu'elles doivent faire, il est très difficile de comprendre comment, et leur analyse constitue un sujet de recherche à part entière (Vondrick et al., 2013; Szegedy et al., 2014; Zeiler and Fergus, 2014).

3.4 *Émotions et mensonges*

Ce qui précède montre que les machines ne souffrent pas de limitations de principe évidentes qui les empêcheraient de développer les mêmes capacités cognitives que les humains.

Néanmoins, on peut se demander si, même si elles pouvaient potentiellement développer les mêmes capacités qu'un humain, elles le feraient. Cette question est différente, car elle fait intervenir le contexte dans lequel la machine évolue. Ses objectifs, son passé, et sa "culture". Deux traits reviennent souvent comme étant inaccessibles aux machines : Les émotions et les mensonges.

Si ces traits ne sont importants dans la discussion que parce qu'ils reflètent des états de conscience forts, et sont utilisés comme des exemples concrets de ce que l'on entend par "conscience de soi", alors on revient au problème abordé en § 2.2. Nous ne pourrons jamais savoir si une machine, ou un autre humain que nous même, qui exprime une émotion ou qui ment, *ressent* effectivement quelque chose de particulier.

Si c'est de l'aspect comportemental dont il s'agit, il n'y a pas d'impossibilité de principe. On peut parfaitement imaginer qu'une machine, par exemple un "robot intelligent", puisse avoir des états fondamentaux correspondant à de la peur, ou de la joie, quand il est dans une situation de danger pour son intégrité physique, ou au contraire quand il atteint un des buts fondamentaux pour lesquels il a été

entraîné (Minsky, 2007). Quant aux mensonges, ils constituent la meilleure stratégie dans le cas où la machine échange des informations et est en compétition avec des tiers pour une ressource limitée. De tels comportements apparaissent naturellement dans des simulations avec des robots (Mitri et al., 2009).

4 CONCLUSION

L'objectif de cet essai est essentiellement de réfuter des arguments de "bon sens" qui sont faux. Grâce à l'extrême complexité des ordinateurs modernes, les machines qu'ils équipent diffèrent qualitativement de celles du quotidien. Elles peuvent en particulier produire des résultats inattendus, et modifier de manière autonome leur fonctionnement selon leurs interactions avec l'environnement.

L'homme est probablement une machine, mais les machines possibles et futures sont bien plus que ce que notre intuition nous dit.

A RÉPONSES À WULFRAM GERSTNER

A.1 *Les machines n'ont pas de capacités introspectives*

Il n'y a pas d'impossibilité de principe à ce qu'une machine soit introspective, c'est à dire qu'elle ait accès à ses représentations internes comme elle a accès à des observations sur le monde extérieur.

A.2 *Une simulation n'est pas la réalité*

Il n'y a pas de séparation claire entre les simulations et la réalité. Il existe par exemple des prothèses auditives ou des rétines artificielles qui remplacent des tissus nerveux. Un traitement de l'information qui se fait normalement dans des neurones est donc effectué à l'aide de calculs dans des microprocesseurs, et aboutit à la même expérience consciente pour celui qui porte la prothèse. Est-ce que ce traitement est une simulation ? Ou bien une réalité physique ?

Wulfram cite l'exemple de simulations informatiques dans lesquelles des molécules d'eau simulées prennent des configurations similaires à de la glace, et il met en avant que cette glace simulée n'est pas *froide*. Cet argument est circulaire dans le débat qui nous occupe : Une intelligence artificielle consciente qui interagirait avec la glace "simulée" ressentirait du froid, donc cette glace serait froide.

A.3 *La chambre chinoise démontre l'insuffisance du test de Turing*

La chambre chinoise est une "expérience de pensées" centrale dans les débats sur la conscience, et sert à démontrer l'insuffisance du test de Turing (Searle, 1980). Dans cette expérience imaginaire un ou plusieurs individus se trouvent dans une pièce fermée, et interagissent avec un interlocuteur extérieur dans une langue qu'ils ne comprennent pas. Ils se réfèrent à des documents qui indiquent de manière formelle, pour chaque phrase qui leur est dite, ce qu'ils doivent répondre.

Il y a deux manières de comprendre cette expérience de pensées, et les deux sont à mon avis peu satisfaisantes.

La première consiste à dire que comme aucun des individus impliqués n'est conscient du contenu de la discussion, mais que cette discussion permet de passer le test de Turing, alors on peut effectivement passer ce test sans état conscient. Cette première interprétation fait une confusion entre l'état de conscience des composants élémentaires d'un système et l'état de conscience du système entier.

De manière similaire, on démontrerait que le cerveau n'est pas conscient parce que les neurones ne peuvent pas l'être individuellement.

La deuxième interprétation de cette expérience est plus généralement qu'un système très simple, et donc évidemment non conscient, peut passer le test de Turing. On pourrait remplacer les individus par un programme informatique primitif qui "regarde dans un livre" la phrase à répondre.

Mais pour des raisons combinatoires ce *livre* ne peut pas exister. La taille de l'univers ne permettrait pas de faire un objet qui contiendrait toutes les discussions nécessaires pour passer le test de Turing. Le contre-argument est alors que le livre peut être remplacé par un dispositif plus complexe pour "compresser l'information", par exemple en regroupant des phrases identiques, ou en utilisant des tables de synonymes. Mais dans ce cas, plus le système devient réalisable, plus il est complexe, et moins l'argument de simplicité tient.

A.4 *Les implications morales sont inacceptables*

Le dernier argument de Wulfram, avancé dans sa duplique, repose sur les conséquences morales de ce débat. Si les humains sont des machines, quelle est la gravité d'un meurtre ?

Le premier contre-argument est que l'univers n'est pas moral, et que les conséquences morales ne révèlent rien sur la réalité physique du monde. Le second contre-argument est que décider quelle est la gravité de la destruction d'une machine est justement la question qui nous occupe : Si les humains sont des machines, alors il existe des machines dont la destruction est un acte moralement extrêmement grave.

B RÉPONSES AU PUBLIC

B.1 *Une machine ne voudrait pas se suicider*

Un argument avancé pendant la séance de questions est qu'un humain, lorsqu'il est soumis à trop de stress et de situations difficiles, peut finalement mettre fin à ses jours, ce qu'une machine ne ferait pas.

Comme pour les émotions et les mensonges, si l'argument sous-jacent est l'expérience consciente, c'est à dire que l'acte suicidaire est un exemple d'état conscient fort, alors on en revient une fois encore au test de Turing. Si l'argument est purement comportemental, on pourrait parfaitement imaginer une machine qui, étant donnés ses objectifs, préfère s'auto-détruire dans certaines situations.

B.2 *Une machine ne peut pas être altruiste*

Une machine peut parfaitement démontrer des principes moraux et altruistes, à nouveau en fonction de ses objectifs. Des simulations avec des robots qui partagent des ressources avec des tiers montrent que de tels comportements apparaissent, et suivent les prédictions de la biologie évolutionniste (Waibel et al., 2011).

B.3 *Une machine est remplaçable*

Finalement, un dernier argument est qu'une machine peut être remplacée. Contrairement à un être humain, on peut lui substituer un autre exemplaire du même modèle si elle est détruite.

Un premier argument serait que le processus de fabrication ou la programmation d'une machine pourrait assurer son unicité à l'aide de paramètres aléatoires qui lui seraient propres. Mais de manière plus fondamentale, même des machines identiques, si elles sont capables d'apprentissage, deviennent uniques après avoir existé assez longtemps. Elles sont le produit de leurs interactions avec l'environnement, qui est propre à chacune.

Comme un animal familier, un robot qui vivrait au quotidien avec un humain, et avec qui il aurait un passé commun, ne serait pas remplaçable par un nouvel exemplaire.

RÉFÉRENCES

D. Chalmers. *The Conscious Mind : In Search of a Fundamental Theory*. New York : Oxford University Press, 1996.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12 :2493–2537, 2011.

D. Dennett. *Consciousness Explained*. Little, Brown and Co., 1991.

D. Dennett. The unimagined preposterousness of zombies. *Journal of Consciousness Studies*, 2(4) :322–325, 1995.

J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, and V. Svetnik. Deep neural nets as a method for quantitative structure–activity relationships. *Journal of Chemical Information and Modeling*, 55(2) :263–274, 2015.

M. Minsky. *The Emotion Machine : Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. Simon & Schuster, 2007.

S. Mitri, D. Floreano, and L. Keller. The Evolution of Information Suppression in Communicating Robots with Conflicting Interests. *Proceedings of the National Academy of Sciences*, 106(37) :15786–15790, 2009.

T. Nagel. What is it like to be a bat ? *The Philosophical Review*, 83(4) :435–450, 1974.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088) :533–536, 1986.

J. R. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3 :417–424, 1980.

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *Proceedings of the International Conference on Learning Representations*, 2014.

A. M. Turing. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society. Second Series*, 42 :230–265, 1936.

A. M. Turing. Computing machinery and intelligence. *MIND : A Quarterly Review of Pyschology and Philosophy*, 59(236) :433–460, 1950.

C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba. Hoggles : Visualizing object detection features. In *Proceedings of the International Conference on Computer Vision*, 2013.

M. Waibel, D. Floreano, and L. Keller. A Quantitative Test of Hamilton's Rule for the Evolution of Altruism. *PLOS Biology*, 9(5) :e1000615, 2011.

M. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision*, 2014.